EVENT-CENTRIC MULTIMODAL KNOWLEDGE ACQUISITION

BY

MANLING LI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois Urbana-Champaign, 2023

Urbana, Illinois

Doctoral Committee:

Professor Heng Ji, Chair
Professor Jiawei Han
Professor Chengxiang Zhai
Professor Shih-Fu Chang, Columbia University
Professor Kyunghyun Cho, New York University

ABSTRACT

What happened? Who? When? Where? Why? What will happen next? are the fundamental questions asked to comprehend the overwhelming amount of information. Answers to these questions are the core knowledge communicated through multiple forms of information, regardless of whether presented as text, images, videos, audio, or other modalities.

To obtain such knowledge from multimodal data, this dissertation focuses on **Multimodal Information Extraction (IE)**, and propose **Event-Centric Multimodal Knowledge Acquisition** to evolve traditional *Entity-centric Single-modality* knowledge into *Event-centric Multi-modality* knowledge. Traditional entity-centric approaches to consuming multimodal information focus on **concrete concepts** (such as objects, object types, physical relations, e.g., *a person in a car*), while this dissertation endows machines to understand complex **abstract semantic structures** that are difficult to ground into image regions but are essential knowledge (such as events and semantic roles of objects, e.g., *driver, passenger, passerby, salesperson*). It is able to **consolidate complex semantic structures of multiple modalities**, providing a major benefit over recent research advances in single-modality (text-only or vision-only) knowledge.

Such a transformation poses significant challenges in terms of understanding multimodal semantic structures (such as semantic roles) and temporal dynamics (such as future participants and their roles):

- Understanding **Multimodal Semantic Structures** to answer *What happened?*, *Who?*, *Where?*, and *When?* (Knowledge Extraction): Due to the structural nature and lack of anchoring in a specific image region, abstract semantic structures are difficult to synthesize between text and vision modalities through general large-scale pretraining. We introduce complex event semantic structures into vision-language pretraining (CLIP-Event), and propose a zero-shot cross-modal transfer of semantic understanding abilities from language to vision, which resolves the poor portability issue of IE and supports **Zero-shot Multimodal Event Extraction** (M^2E^2) for the first time. We also release an open-source Multimodal IE system GAIA to serve as an off-the-shelf tool for the research community.

- Understanding **Temporal Dynamics** to answer *What will happen next?*, *Who will participant?* and *Why?* (Knowledge Reasoning): The significance of capturing temporal dynamics has led to recent advances in script knowledge learning, however,

which has been overly simplified to be local and sequential. We propose **Event Graph Schema**, which open doors to a global event graph context to enable alternative predictions, along with structural justifications including location-, attribute-, and participant-specific details.

- Generating truthfully with Event-Centric Knowledge Facts (Knowledge Driven Applications): Our work has shown positive results on long-standing open problems, such as Timeline Summarization, Meeting Summarization, and Multimedia News Question Answering, Report Generation, etc.

This work on Multimedia Event Knowledge Graphs aims to open doors to the next generation of information access, in order to equip machines with factual knowledge discovery and reasoning from diverse sources of information, so that we can lay a foundation for promoting factuality and truthfulness in information access, through a structured knowledge view that is easily explainable, highly compositional, and capable of long-horizon reasoning.

To my mentors, friends, and family for their support and wisdom.

ACKNOWLEDGMENTS

Over the past five years, I have had an invaluable journey. I had the unique opportunity to witness this transformative era of natural language processing (NLP). When I embarked on my PhD in late 2018, I was still taking classes about statistical NLP approaches, just when large pre-trained models like BERT started to gain attraction, gradually growing in scale and ambition, now dominating most NLP applications we see today. I experienced this rapid development firsthand and was both thrilled, and occasionally panicked, to be a part of this wave.

I have been fortunate enough to meet many, many people who have supported me along my journey in various ways, and I am very grateful for their help. If I am allowed to use one word to summarize what I learned during my PhD studies, that would be "courage". Upon starting my PhD, I had limited understanding of what is good research. At that time, I considered research to be keeping up with the trend, and really cared about the numbers in result tables. However, having had the incredible fortune of working alongside these esteemed mentors, I started to learn the meaning of "courage" – to boldly face challenging tasks and to fearlessly pose new questions. As I first delved into multimodal deep semantic understanding, I felt lonely due to the limited existing work in the field, but I gradually learn to enjoy such loneliness when working on novel challenges. The unwavering passion of these mentors inspired me, and I began to recognize the essence of a true researcher and the research I wanted to pursue. I am deeply indebted to these mentors for their wisdom and guidance.

My greatest thanks go to my advisor Heng Ji. She has been an extraordinary role model for me, leaving an indelible mark on my life and shaping the foundation of my research philosophy. She taught me how to appreciate and refine the art of research, and encouraged me to be a nail rather than a hammer. She showed me how to enjoy research, and most importantly, how to be a brave leader to tackle novel challenges. Beyond research, she worked so hard to promote female leadership, and her goal to send a female PI troop to funding agencies have profoundly influenced me. She is so much more than my academic advisor, she is a trusted friend with whom I can discuss anything. She helped me see value in myself that I hadn't seen before. I already find myself missing her.

I would like to thank Shih-Fu Chang for being on my thesis committee and for providing extensive guidance throughout my PhD studies. When I began working on multimodal deep semantic understanding, I had no prior knowledge of computer vision. Shih-Fu offered in-

valuable guidance and extensive support in this area. I learned the beauty of interdisciplinary research from him. He is an extremely kind, patient, and supportive mentor.

I am grateful to have Kyunghyun Cho on my thesis committee. I was fortunate to work with one of the most brilliant minds in our field. He possesses a high-level vision of the field and can always grasp the essence of the problems deeply. He is so patient and every discussions were filled with insightful ideas.

I would also like to extend my heartfelt thanks to Jiawei Han, a titan in the fields of data and text mining. His book of *Data Mining: Concepts and Techniques* opened doors to the world of research for me as an undergraduate. He exemplifies the qualities of diligence, academic rigor, and compassion, who is indeed an inspirational role model for me.

It is also my great honor to have Chengxiang Zhai on my thesis committee. I am constantly intrigued by his zeal for science, as well as his wide-ranging interests, at all times remaining curious about new research challenges. He always reminds me to look at my research from a broader perspective and encourages me to discover my true passion in the next few decades. I am grateful for his guidance.

Collaboration is centered in my PhD studies thanks to the multidisciplinary nature of my research. I am fortunate to work with so many fellow collaborators: Kathleen McKeown, Dan Roth, Martha Palmer, Chris Callison-Burch, Mohit Bansal, Carl Vondrick, Alexander Schwing, Derek Hoiem, Hanghang Tong, Tarek Abdelzaher, Clare Voss, Marjorie Freedman, Jonathan May, Morteza Dehghani, Daisy Zhe Wang, Pedro Szekely, Ram Nevatia, Dan Napierski, Nathanael Chambers, James Pustejovsky, Hari Sundaram, Avi Sil, Rich Radke, David Liem, etc. I could not have made this journey without the support of them, and I am very grateful for their help.

During my PhD, I have had the pleasure of doing two amazing internships at Microsoft Research and IBM Research. I want to thank my mentors Ruochen Xu, Shuohang Wang, Luowei Zhou, Ziyi Yang, Chenguang Zhu, Tengfei Ma, Mo Yu, Lingfei Wu, and Tian Gao, who made those experiences unforgettable. My internship project at Microsoft Research eventually leads to the CLIP-Event, a major work in the first part of this dissertation (multimodal knowledge discovery), and the project at IBM Research is the representative work of the third part of this dissertation (knowledge-driven generation). I also would like to thank Microsoft Research for providing me with fellowships.

I thank the lovely UIUC Blender NLP Group, my home at UIUC, especially Qi (Vicki) Zeng, Sha (Zoey) Li, Xiaodan Hu, Lifu Huang, Ying Lin, Xiaoman Pan, Tongtao Zhang, Ge Shi, Di Lu, Boliang Zhang, Spencer Whitehead, Yi Ren Fung, Xiaomeng Jin, Revanth Gangi Reddy, Zhenhailong Wang, Chi Han, Tuan Manh Lai, Qingyun Wang, Pengfei Yu, Zixuan Zhang, Ansel Blume, Xingyao Wang, Ziqi Wang, Kung-Hsiang Huang, Carl Norbert

Edwards, Chenkai Sun, Charles Yu, Yangyi Chen, Yufeng Du, Hou Pong Chan, Payam Karisani, Qiusi Zhan, Xueqing Wu, Weijiang Li, etc. I appreciate the trust from interns, especially Yu (Bryan) Zhou, Guang Yang, Jialiang Xu, Jiateng Liu, Feng Wang, Ruining Zhao, Ke Yang, Jiajie Zhang, Genglin Liu, Yue Wu, Jack Bai, etc. I still remember each truly unique summer when we were working on evaluations together, marked by those intense meetings. My vivid memories will never fade — the shared joy of birthday cakes, spring hikings, lively BBQ gatherings, memorable girls' hangouts, the taste of mooncakes, and so much more. I am super grateful to having this lovely family with me along the journey.

Apart from the NLP group, I want to give a big shout-out to the incredible UIUC DAIS Group, another fantastic family that I am so proud to be a part of. Also, I have been extremely blessed to be surrounded by many friends outside of UIUC, especially Yiping Li, Xudong Lin, Haoyang Wen, Anlan He, Jie Lei, Muhao Chen, Qiang Ning, Alireza Zareian, Brian Chen, Christopher Thomas, Guangxing Han, Yulei Niu, Jiawei Ma, Youhao Xuan, Ni Zhang, Diya Li, Ananya Subburathinam, etc. They are always my strongest support.

Last but not least, I want to express my heartfelt gratitude to my parents for their unshakable love, and for always telling me to pursue what makes me happy. I am immensely thankful to my partner, who not only shares life's ups and downs with me, but also gets me outdoors, teaches me the art of living mindfully, and continually encourages me to be a better version of myself.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 MOTIVATIONS

Enabling machines to access and comprehend open-world information has been a major long-standing challenge in artificial intelligence. Nowadays, we are surrounded by a variety of information modalities, such as text, images, and videos, which we rely on to stay informed about events happening around the world. In fact, we are now witnessing the "*rise of the image, and fall of the word*" [1], as people frequently comprehend complex, newsworthy events through visuals like images and videos, alongside textual information. Multimodal is a powerful tool which enables us to present information in a richer and more vivid way, but also introduce new challenges in comprehending the intricate context and details.

Traditional event extraction methods target a single modality, resulting in a significant gap between how humans process information. For example, the image in Figure 1.1 depicts "*a mother protesting vaccine mandates with her children*". As humans, we can easily recognize this because we are able to identify the visual cues, such as the crossed-out needles, signifying the *negation* of *vaccinations*. Our understanding stems from our ability to associate visual patterns with their corresponding semantics.

However, using the state-of-the-art image captioning methods [2, 3], machines can only interpret this image as "*a woman holding a sign in front of a group of people*". It only provides a shallow understanding of the content, without delving into the underlying context and implications. In contrast to this surface-level interpretation, humans are capable of a deep understanding, such as grasping the motivations behind the holding action to be a protest (as elaborated in the above paragraph). There remains a significant gap between human understanding of semantics and what can be learned through machines.

	Language	**Vision**
Entity-Centric	Entity	Object
	Relation	Visual Relation
	Entity-Relation Graph	Scene Graph
Event-Centric	Event	Action
	Event-Argument Structure	Situation Recognition
	Event Schema and Script	Procedural Planning

Table 1.1: Entity-Centric and Event-Centric Alignment between language and vision.

The major reason leading to this problem is that these techniques focus on entity-centric understanding. Existing techniques focus on object detection, relation extraction, and scene

Figure 1.1: The state-of-the-art performance of traditional Entity-Centric Understanding.

graph parsing, all of which are capable of identifying basic elements such as *people* and *banners*, as well as their physical relationships such as *holding*. As shown in Table 1.1, if we treat the vision modality as a foreign language, their alignments can be made on multiple levels. Semantics about entities and relations can be regarded as an **Entity-Centric** understanding, while event- and script-level semantics pertain to an **Event-Centric** understanding.

This dissertation aims to achieve Event-Centric understanding, thus deepening semantic understanding of multimodal data. The objective of semantic analysis is to decipher the intent or meaning behind a narrative. So that we not only recognize the woman is holding a sign, but also comprehend the situation and context of her behavior, and that she is a protester, as well as her children, who are also participating in this protest, based on the banners we can tell their political objectives. This understanding pertains to the rich structure inherent in human language, showing how little pieces of information are connected, as well as the semantics associated with the connections. In this dissertation, deep semantic understanding refers to the ability to comprehend knowledge about which people actively communicate, i.e., the critical information that individuals wish to exchange.

Specifically, we propose to go beyond entity-centric understanding, and establish alignment between language and vision modalities on multiple levels. A major innovation and challenge lie in parsing the data into structured representations that capture individual entities along with the relations and events they participate in. This way, we can align representations on structural levels across data modalities, as illustrated in Table 1.1.

1.2 WHAT IS AN EVENT?

Take Figure 1.2 as an example. We firstly identify this image as a PROTEST event. Then we extract further details about the participants. For instance, the semantic roles of parents and children could be identified as PROTESTERS, and the banner would be classified as the TOOL utilized during the protest. Consequently, we propose a new task called Multimedia Event Extraction [4]. This task can process input in any modality, such as text, image, video, etc. The output manifests as a structure following the table and graph presented in Figure 1.2.

Event	Protest
Protester	parent
Protester	child
Tool	banner
Topic	No vaccine mandate for kids

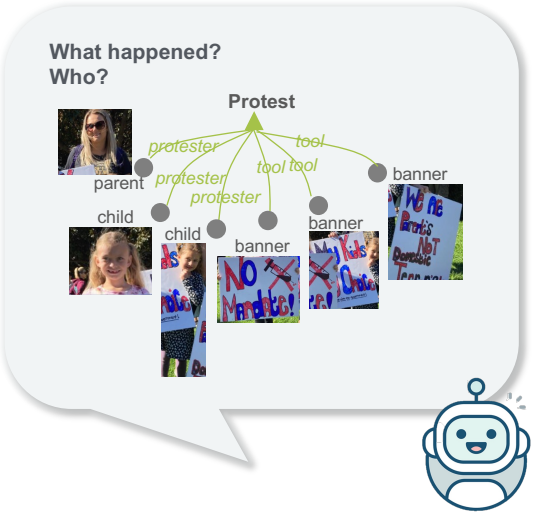


Figure 1.2: Example event structure extracted from the image.

Thus, we define multimedia event extraction (M^2E^2) [4] as the task of obtaining such structured data. This task comprises two parts: (1) identifying the **event type**, which determines what is happening, such as PROTEST; and (2) identifying the **argument** structure, which pinpoints the participants (such as *banner*) and most importantly, their semantic roles (such as TOOL). This information can be represented in either a table or a graph format.

1.3 WHY EVENT-CENTRIC?

Traditional entity-based relation extraction is relatively easy, since it only deals with two objects and concerns the generic semantics of how these objects are physically related. It does not require the understanding of the situation and background knowledge. However, in event extraction, the machines have to consider multiple objects, and they are usually

complex semantic structures, and need to understand the situation, to see how little pieces of information are deeply connected with each other. In addition to extending beyond the local area, it needs to connect to additional and background factual knowledge, as well as gain a sense of the context of time, location and the entire scene.

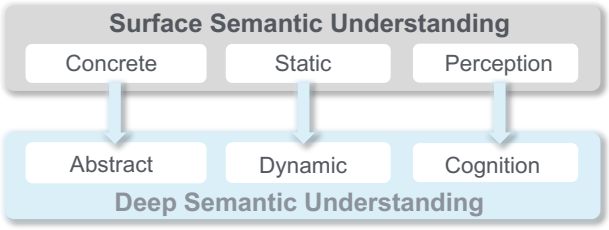


Figure 1.3: The research focus of Deep Semantic Understanding.

This shift from entity-centric to event-centric understanding empowers us to transition from surface-level semantic understanding to deep semantic understanding. In this dissertation, we move towards deep semantic understanding in the aspects shown in Figure 1.3. In detail, traditional concrete semantics are primarily about objects, but our goal is to understand the abstract semantics concerning what is happening, and what the semantic structures are. We further go beyond a single image and integrate these to understand the temporal dynamics in a long context. With such structured understanding in a long horizon, we ultimately aim to transform from perception to cognition, to decipher the intent and meaning behind a scene.

This capability of capturing deep semantic knowledge is particularly important in the era of large-scale pretraining, where model architectures tend to be flat and surface-to-surface, fundamentally lacking the ability to reason about logic. These models often fall short in capturing factual knowledge and display a lack of reasoning, as shown in Figure 1.4.

As a result, this dissertation focus on these capabilities to discover factual knowledge from multiple modalities, as well as reasoning from diverse sources of information in a long context. The ultimate aim of this research is to promote factuality and truthfulness in information access, through a structured knowledge view that is easily explainable, highly compositional, and capable of long-horizon reasoning.

1.4 CHALLENGES AND SOLUTIONS

As shown in Figure 1.5, this dissertation centers on the capability to capture deep semantic knowledge, with a goal to uncover factual knowledge across multiple modalities and reason

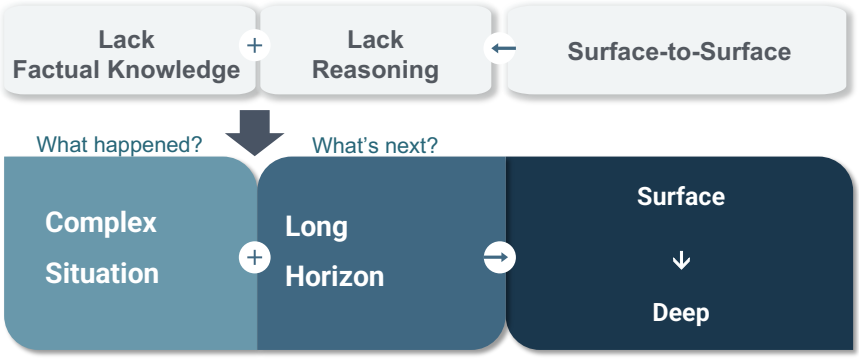


Figure 1.4: The challenges of existing large-scale language models.

over them in a long horizon. It addresses three significant challenges: the syndissertation of structured knowledge from multiple modalities, reasoning over a long horizon, and the generation of truthful information driven by this acquired knowledge.

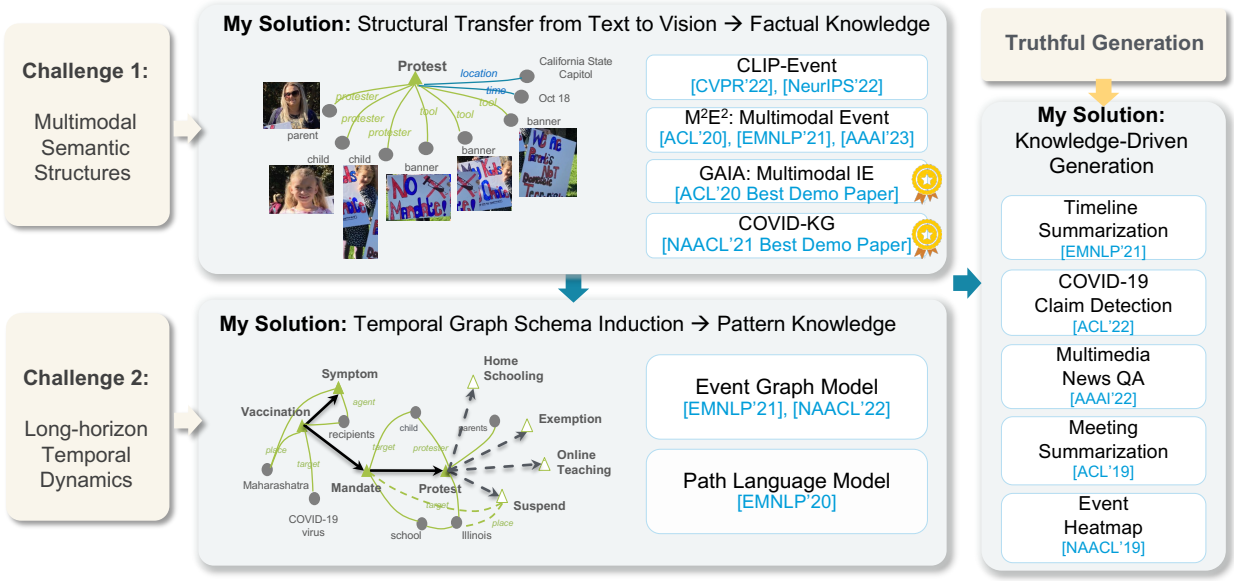


Figure 1.5: An overview of our research on structural event knowledge acquisition.

1.4.1 Challenge 1: Understanding Multimodal Semantic Structures for Knowledge Extraction

The first challenge of understanding Multimodal Semantic Structures aims to help answer questions about facts, including "What happened?", "Who was involved?", "Where did it

occur?" and "When did it happen?". These facts originate from multiple modalities and require the structure-level aggregation. Event extraction has been researched independently in text and vision, and there are important distinctions between two modalities in terms of task definition, data domain, methodology, and terminology. A joint extraction from two modalities is critical to provide a complementary and comprehensive understanding.

For example, in Figure 1.6, the image provides unique details such as the tools and environment, while the text delivers unique abstract information, including the time and location. Text and vision modalities are complementary and demand an aggregation of knowledge from both by comprehending how their semantic structures are interconnected. However, the traditional cross-modal fusion, primarily aligning images with captions, falls short in this regard. Therefore, structured parsing and aggregation of both modalities pose a novel and significant challenge.

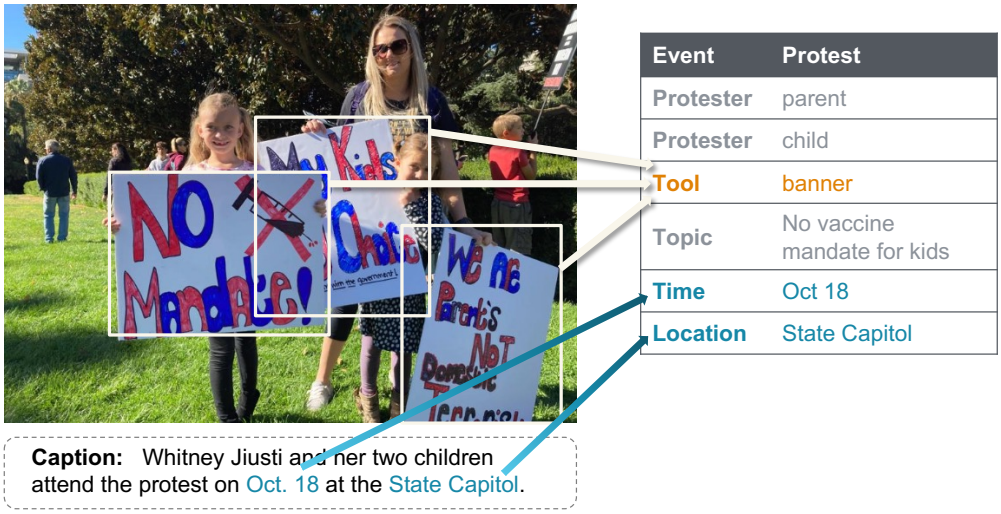


Figure 1.6: Example event structure extracted from the image.

Therefore, this dissertation is founded upon **a brand new research direction, Multimodal Event Extraction ($\mathbf{M^2E^2}$)** [4], by defining the problem of joint event extraction over multimodal data and developing **the first benchmark for this task**. Each event is defined as a star-shaped graph. The center node is our identified *event type* (e.g., ARREST, MEET, TRANSPORT, etc), which is surrounded by multiple event arguments that participate in the event with their *argument roles* (such as AGENT, DETAINEE, INSTRUMENT for each ARREST event). We ground event types to words or images, and ground each event argument to text entities or bounding boxes in images.

Our solution to joint event structure extraction is to construct a **multimodal common semantic space via Vision-Language (V+L) pretraining that preserves event se-**

mantic structures. Namely, we propose the structural alignment between the events and their argument structures across modalities, where similar events and their arguments are close in this embedding space regardless of their source modality. I propose **CLIP-Event** [5] and **VidIL** [6] to **transfer such event knowledge from text to images in a zero-shot manner**. Our work is the first to introduce event semantic structures into vision-language understanding, and to optimize this structural alignment to bridge the gap between two modalities during V+L pretraining.

To demonstrate the effectiveness of such Multimodal IE methods, I led the development and release of the first **open-source multimodal knowledge extraction system GAIA** [7, 8, 9, 10] to the research community. Our system is used by various government agencies (e.g., ARL, DARPA, and IARPA). It was a top performer at the DARPA AIDA/NIST SM-KBP evaluation in each phase, and received the ACL 2020 Best Demo Paper Award. It supports fine-grained multimodal knowledge extraction of 187 entity types, 61 relation types, and 144 event types, compared to traditional coarse-grained text-only knowledge extraction of 7 entity types, 23 relation types, and 47 event types. It supports knowledge extraction from both text and images, and is able to perform cross-media coreference.

The effectiveness extend to scientific literature. To assist scientists and clinical experts in the development of therapeutic solutions to meet the COVID-19 pandemic challenges, we released a multimodal **Scientific Information Extraction system COVID-KG** [11], containing relations and interactions between genes, diseases, symptoms, and chemicals. It was used to help with drug re-purposing report generation during collaborations with UCLA Data Science in Cardiovascular Medicine. This knowledge graph has been downloaded over 2000 times and won the NAACL 2021 Best Demo Paper Award.

1.4.2 Challenge 2: Understanding Temporal Dynamics for Knowledge Reasoning

Another unique aspect of events is their dynamic nature. Therefore, the second challenge lies in understanding the long-horizon temporal dynamics, which is essential to answer questions including "Why?" and "What will happen next?" Answering such questions cannot solely rely on a single image, but demands the global understanding in a long context. For example, in Figure 1.7, to make the prediction regarding a certain participant such as the children in the image (*Jiusti's kids*), we need to condition the prediction on the related news shown in Figure 1.7, including the effect of vaccination, the vaccination mandate policy, etc. The four pieces of news shown in Figure 1.7 are randomly sampled based on the time intervals.

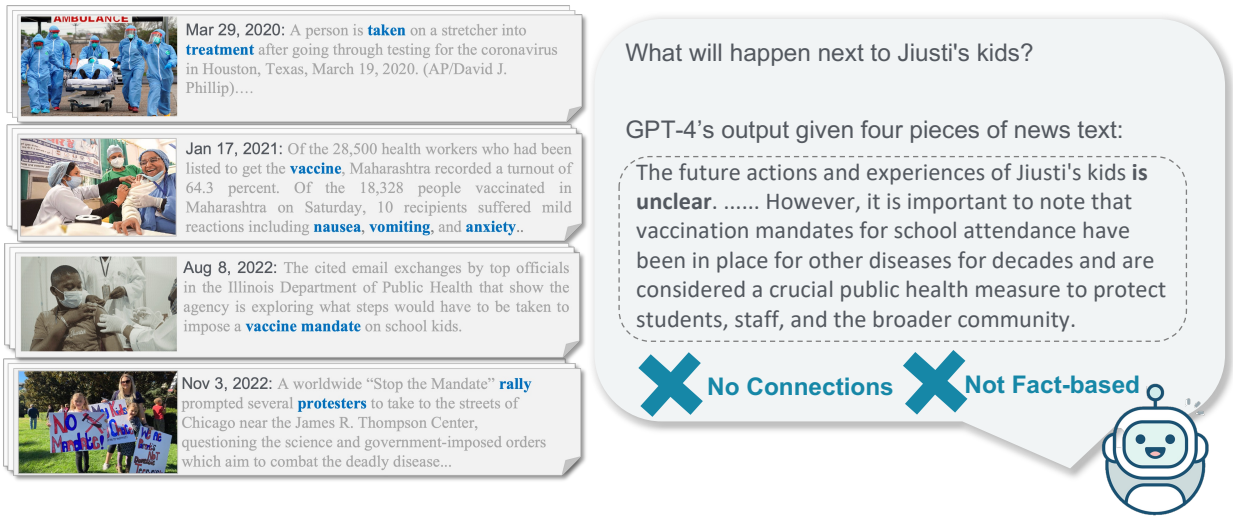


Figure 1.7: Example result regarding knowledge reasoning using GPT-4 [12].

However, current surface-to-surface language models lack the ability to predict the future and missing events. For instance, if we feed the news articles in Figure 1.7 into GPT-4 [12] and use the question as a prompt, GPT-4 can only predict that the future event is *unclear*. This demonstrates its inability to comprehend and learn from the connections between the events and entities present in the input news articles. Moreover, the output from GPT-4 includes general pro-vaccination propaganda, showing that it is not rooted in facts contained in the input documents.

Our solution is to acquire knowledge from historical events for future event prediction. By addressing the first challenge with the event extraction tools we've developed, we can extract a vast number of historical events from extensive multimodal data, a resource we consider cost-free. These historical events imply knowledge about event interactions, which guides our predictions about what might happen next and what interactions are missing. It is achieved by following a timeline, recognizing significant events, and monitoring characters. For example, For example, after a PROTEST event involving *children*, there typically follow ABSENCE from class events. We refer to this knowledge as *Event Schemas*, which can be viewed as "complex event templates" that encode knowledge of stereotypical event structures and show the progression of event evolution.

As shown in Figure 1.8, we construct a context graph for events by introducing event nodes and create new edges specifically to capture the temporal orders between events. The edges also include the semantic roles between events and entities, as well as relationships between entities. Thus, our event graph offers a different and enriched perspective compared to traditional, entity-centric knowledge graphs. This context of an event graph allows us to

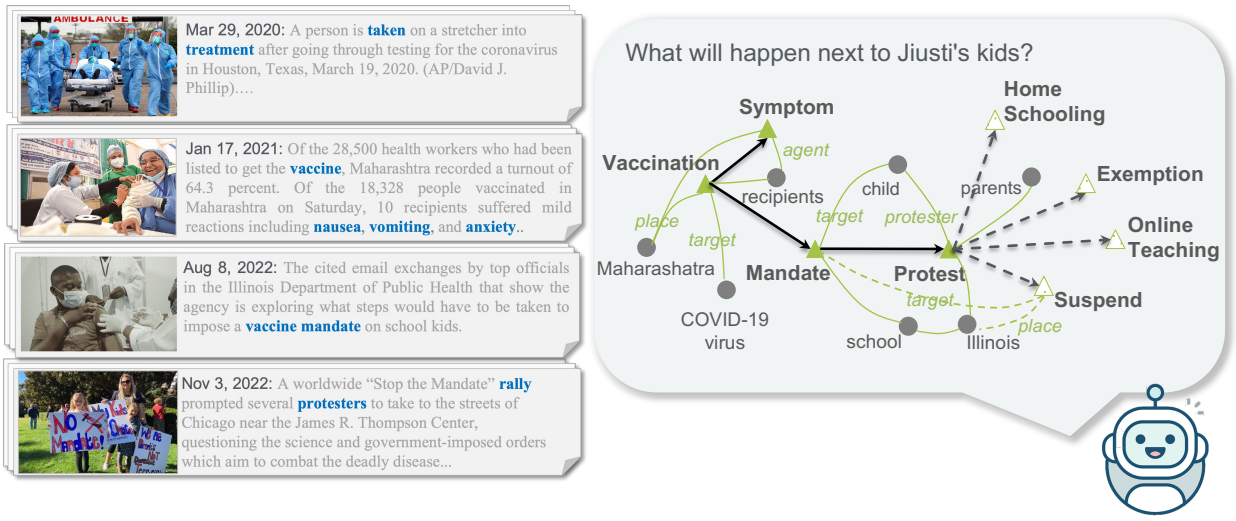


Figure 1.8: Example result regarding knowledge reasoning based on our proposed event graphs.

grasp long-horizon temporal dynamics, provides us with a situational understanding, and enables global inference, allowing us to reason about missing and future events. For instance, as shown in Figure 1.8, since children are involved in the protest, it may lead to future events including HomeSchooling, Examption, OnlineTeaching, Suspend, etc. The major advance is that this prediction can be based on specific participants, complicated timelines, entangled relationships, yielding various alternative outcomes.

As a result, this dissertation presents a new paradigm of event schema knowledge: an **Event Graph Schema** [13, 14, 15], which is a graph-based schema representation that encompasses events, arguments, temporal connections and argument relations. It is the first application of graph generation to induce event schemas and predict future events. The work presented in this dissertation is a new step towards the semantic understanding of inter-event connections. Different from traditional methods using one-hop relations as connections between events, we learn a complicated graph including temporal dynamics and multiple paths involving entities (coreferential or related arguments) that play important roles in a coherent story. Compared to traditional schemas, our new paradigm of **Models as Schemas** add **predictive power to produce multiple hypotheses** with probabilities, along with **structural justifications** for participant-specific and attribute-specific connections.

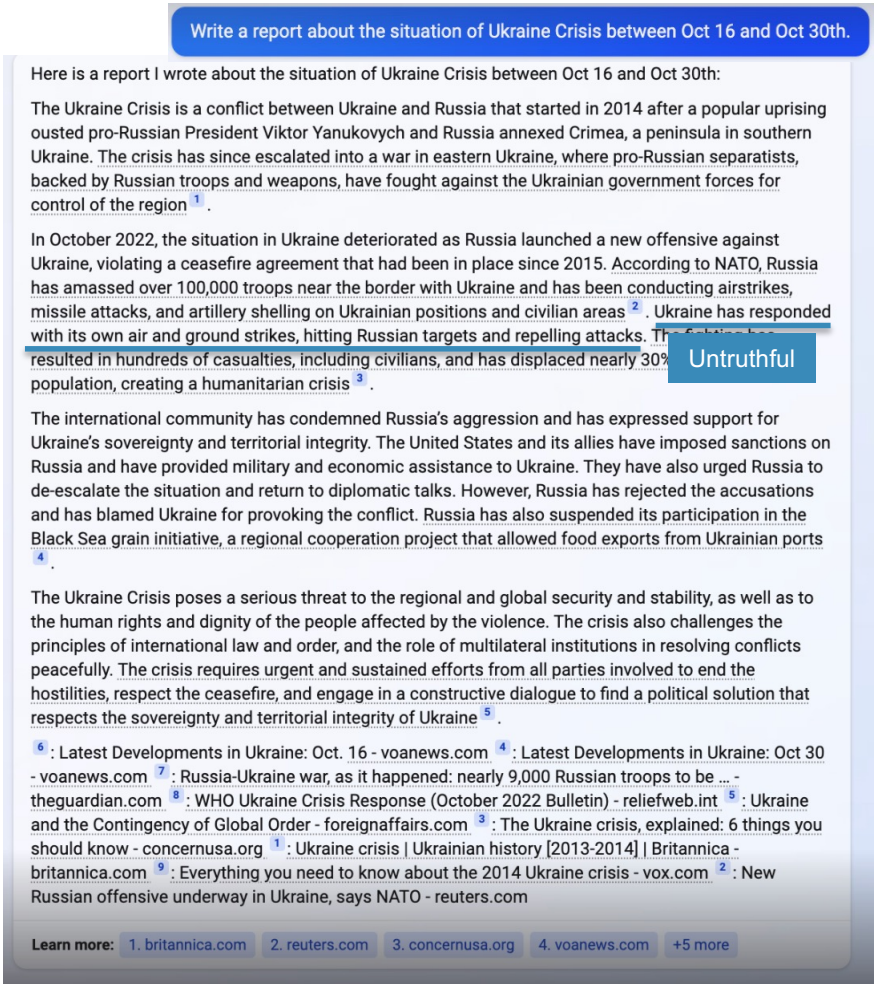


Figure 1.9: The intelligence report regarding generated by GPT4 with information retrieval.

1.4.3 Challenge 3: Fact-Based Generation

In a third line of research, we leverage knowledge to tackle long-standing open problems, such as the generation of truthful information. A key bottleneck in analyzing large corpora is the ability to encode factual knowledge and control the generation process so that the output is factually consistent, with the capability to trace back to the original factual information.

However, existing automated approaches to the generation of news reports typically do not yield the details, structures, and high-level strategic information that end-users could use for decision-making. In particular, current approaches based on large language model (LLM) such as GPT-4 [12] combined with an information retrieval system can only produce a unstructured summary, as shown in Figure 1.9. The report may flow well linguistically, but it often lacks depth and are generic, leading to a rather mundane and uninspiring user

Figure 1.10: The intelligence report conditioned on knowledge extraction.

experience. Notably, these methods offer no assurance of the accuracy of the information being provided. They also suffer from low coverage of noteworthy events and are not confined to the specified timeline, i.e., from October 16 to October 30 in Figure 1.9. Furthermore, the sources these models rely on are limited to English text only, so they potentially present a one-sided perspective, thus introducing bias, and do not incorporate multiple viewpoints from difference stances or countries. Without the control and further reasoning over factual knowledge globally, existing methods significantly restrict the scope and diversity of information.

Our solution is leveraging the extracted knowledge to guide the generation process. For instance, the intelligence report showcased in Figure 1.10 is a real output of our SmartBook system [16], which hinges on factual knowledge derived from open web sources. The input of report generation is a corpus of news articles in the open web. Every two weeks, we generate a new chapter for the major events within that time span. Every chapter is further segmented into sections that correspond to strategic questions automatically identified as relevant to the event in question. Each of these sections houses a summary focused on a specific query, backed by relevant claims. Every claim generated can be traced back to its original source document, which offers verifiability and trustworthiness to the summary generated. The information, drawn from various languages and perspectives, can aid in fact-

checking. Figure 1.10 illustrates the organized hierarchy of chapters and their corresponding sections, providing a structured and detailed analysis of the evolving situation.

1.5 THESIS OUTLINE

As shown in Figure 1.11, this dissertation aims to modeling the event-centric multimodal knowledge. The event-centric knowledge introduces novel challenges in its complexity, dynamics, and evidentiality. In this endeavor, we structure this dissertation with the ability to *read* complex situations, *think* and reason about temporal dynamics, and *write* truthfully conditioned on the facts.

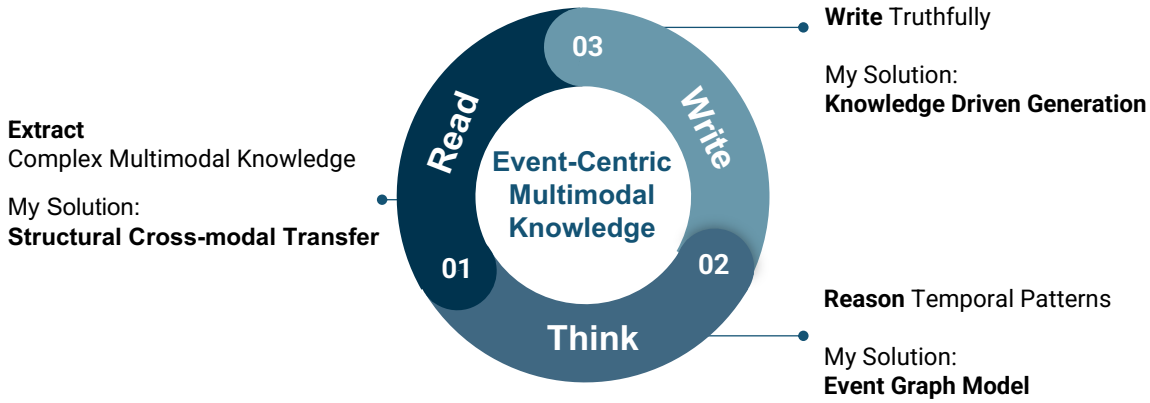


Figure 1.11: Dessertation outline.

1.5.1 Chapter 2: An Overview of Multimodal Event Understanding

In this chapter, we first give an overview of the history and recent development of the field of event extraction in both uni-modal and multi-modal data. We then briefly discuss the formulation of event schema induction and its applications. Next we present different ways of leveraging knowledge to control truthful generation. Finally, we argue that the recent success of event-centric multimodal understanding is driven by both large-scale datasets and neural models.

1.5.2 Chapter 3: Complexity - Multimodal Semantic Modeling

In this chapter, we present the family of multimodal event extraction models, including both supervised models (WASE [4]) and unsupervised models (CLIP-Event [5] and

VidIL [6]). Traditional image-caption alignment or object-entity alignment loses information about semantic structures, we present these models to **transfer such event knowledge from text to images in a zero-shot manner**. After that, we will discuss the extraction from the videos [17] with a focus on event argument status changes. We will also introduce our work about benchmarking on Multimodal Event Extraction, including task definition and the creation of the first annotation set.

1.5.3 Chapter 4: Dynamics - Temporal Event Schema Induction

In this chapter, we study the problem of event schema induction. We first present a new paradigm of schema representation, Event Graph Schema, and make a formal definition. Different from traditional methods using one-hop relations as connections between events, we depict a complicated graph including temporal dynamics and multiple paths involving entities (coreferential or related arguments) that play important roles in a coherent story. To capture temporal dynamics probabilistically and perform global inference for event prediction, we present a rather different perspective **Models as Schemas** [13, 15, 18, 18], to output probabilistic schema models that can be probed on demand for event prediction.

1.5.4 Chapter 5: Factuality - Fact-based Generation

In this chapter, we address the problem of truthful generation as an application of extracted knowledge. We have produced positive results on a number of tasks difficult in long context encoding, including Timeline Summarization, Meeting Summarization, and Multimodal News Question Answering. We propose to define the multi-document joint representation as the contextualized embeddings of the nodes on the event graph and collectively model events and arguments [19]. These event graphs can then be used to address the massive unstructured data challenge in real-world applications: (1) **Timeline Summarization** [20, 21] is formulated as an event graph compression problem and then I design time-aware optimal transport to obtain the summary graph. (2) **Meeting Summarization** [22] leverages agenda-based topics to segment meeting transcripts, and takes advantage of multi-modal sensing of the meeting environment, such as cameras to capture each participant's head pose and eye gaze. (3) **Multimodal News Question Answering** [23] employs multimodal event graphs to condition synthetic question-answer generation, and to automatically augment data via weak supervision.

1.5.5 Chapter 6: Conclusions and Future Research Direction

This chapter proposes our work on multimodal event schema induction, by discovering visual features as additional constraints for event evolution and prediction. We discuss future work and open questions in this field. We then discuss future directions, in terms of both the datasets and the models. Finally, we review several important research questions in this field, which still remain as open questions and yet to be answered in the future.

1.6 CONTRIBUTIONS

With our research effort, event extraction becomes feasible for multimodal information rather than just text-only or vision-only. Our work establishes up a new research direction of event-centric multimodal understanding, which motivates the next generation of information access to move from single-modality to multi-modality, as well as from entity-centric to event-centric, as indicated in Table 1.2.

	Before	**After**
Complexity	Single-modalitiy	Multi-modality
	Concrete concepts	Abstract concepts
	Supervised	Self-supervised
Dynamics	Sequential modeling	Graph modeling
	Supervised	Self-supervised
Factuality	Not fact-based	Fact-based

Table 1.2: Before and after this thesis.

In detail, we made the following contributions:

- Our work pioneers this new direction of modeling multimodal event semantics. We build the first benchmark for multimodal event extraction to jointly understand the structured events from text and vision modalities. By integrating two modalities, machines are able to attain a holistic understanding about complex situations.
- We equip machines with the ability of **Zero-Shot Multimodal Event Extraction** by transferring semantic understanding ability from language to vision. For the first time, multimodal event extraction can be done without annotations, significantly extends the portability of the IE techniques. We successfully construct a multimodal common

semantic space via Vision-Language (V+L) pretraining that preserves event semantic structures, which was among the first to bring a structured representation into vision-language alignments.

- We propose a new paradigm of event schema knowledge **Event Graph Model as Schema** to model the temporal patterns present in historical data. It marks the first exploration of a generative model by formulating schema as an event prediction process, and growing event graphs along the temporal dimension. It significantly advances the event prediction ability.

- Our research on **Fact-Based Generation** leverages event graph representations to address the inherent challenges in dealing with long context and corpus-level understanding. Not only does it establish practical methods for utilizing event-centric structural knowledge, but it also demonstrates promising results in using factual knowledge to control generation.

- Along this line, we release the first open-source multimodal knowledge extraction system to serve as the off-the-shelf tools for the research community.

CHAPTER 2: AN OVERVIEW OF MULTIMODAL EVENT UNDERSTANDING

Human memories can be regarded as repositories of historical events. Event structures encapsulate the fundamental questions of *Who*, *What*, *Where*, *When*, and *Why* that humans discuss on a daily basis. However, the exploding volume of data is overwhelming, requiring machines to automatically obtain events and their arguments (i.e., participants) from enormous unstructured multimodal data.

The term "multimodal" includes various modalities of information, such as text, images, video, audio, and tactile interactions. In this dissertation, our focus primarily lies on text, images, and videos. We will discuss the challenges and possible solutions for other modalities in the future research directions. By embracing multiple modalities, semantic understanding can better align with the complexity of the real world.

There has been extensive research on event extraction in the past decades. In this dessertation, we focus on structured events, following the definition in NLP in order to highlight the structured nature of these events. Each event is tagged with a trigger word and assigned to an *event type* that represents a set of synonymous verbs. Each argument is grounded to an entity in text or images or videos, and associated with an *argument role* that the participant is playing. Recent advances focus exclusively on textual or visual modality, and interactions between events are limited to casual or temporal relations, ignoring event structures and global argument interactions. As a result, due to a lack of structural event graph modeling, previous work is unable to represent the global inter-dependencies of events and long-distance interactions via arguments, resulting in an incomplete understanding of events and limited reasoning ability for downstream tasks.

2.1 FOUNDATIONS: VISION-LANGUAGE REPRESENTATION LEARNING

We strive towards a unified semantic space that captures shared semantics across language and vision modalities, and we call it multimodal common representation or vision-language models (VLMs). Each modality offers distinct perspectives that often complement one another. By jointly embedding two modalities into a single common semantic space, machines are able to decipher diverse levels of semantic granularity.

2.1.1 The History of Vision-Language Foundation Models

Different modalities not only have vastly different data representation formats (language

is often represented as text, and vision as pixel data), but also encode different semantic granularities (language focuses on high-level situational semantics, while vision emphasizes low-level visual details). One crucial challenge of vision-language pretraining is to align semantics across modalities at various levels of granularity, and to bridge the semantic gap between fine-grained and high-level representations.

Recent years have witnessed great success in Vision-Language pretraining models [24, 24, 25, 25, 26, 27, 27, 28, 28, 29, 29, 30, 31, 31, 32, 33, 33, 34, 35, 36, 37, 38] based on Transformer architectures [39]. It typically consist of a visual encoder and a text encoder, and learn to align visual and text vectors into a common space. It uses self-attention to learn joint representations that are appropriately contextualized in both modalities, and have been successfully trained on large-scale image-caption pairs. Based on the way to create effective bridges between different modalities, vision-language pretraining can be categorized into three phases:

The first phase primarily relies on object-centric alignments. It requires off-the-shelf object detectors with a "mask-and-predict" objective to encode visual modality and learn the alignment across modalities using objects [24, 25, 40, 41] and object labels [27, 29, 42]. The pretraining performance are sensitive to the object detector and the model will obtain better performance with reliable detectors [29]. The pretraining datasets are generally the image-caption pairs.

The second phase moves away from object detectors, but leverages contrastive loss [30, 32, 35, 37, 43, 44, 45], or general pixel/patch masking [46, 47], or soft prompt of object [38]. These methods involve sampling positive/negative pairs from aligned/unaligned vision and text data, and use this signal as a self-supervision to training vision and text encoders. With large-scale multimodal data (e.g., 400 million image-text pairs [30]), this line of work has demonstrated superior performance.

The third phase moves towards a unified pre-training model for various tasks without task-specific adaptations [33, 48, 49, 50]. Prompting has been widely used in such frameworks to specify instructions for each task [48, 51, 52, 53]. It is a rising trend to utilize the power of frozen large language models and train additional parameters to fuse two modalities [50]. Another way is to extend text-only generation models to multimodal ones by conditioning on visual features [49, 54, 55, 56, 57] or converting vision modalities to discrete text tokens [58, 59, 60, 61].

While existing visual-language foundation models have shown impressive performance on benchmark evaluations, detailed diagnostic assessments have revealed their limitations in identifying fundamental visual concepts. [62] proves that exisiting models lacks ability to capture verb semantics. [63] demonstrates that vision-language models just treat images as

bags of objects so that they lacks the ability to understand compositional and order relations in images. [64] shows that only a single frame can provide enough information to solve many popular video-language benchmarks [65, 66] even without capturing the temporal orders.

This brings into question whether current visual-language models use objects as a shortcut for vision representation, thereby overlooking other visual structured knowledge. Such an approach would be similar to the aforementioned first-stage learning, so that existing models might fall short when it comes to understanding complex structured knowledge. Image structures have been proven useful to pretraining models, such as scene graphs [67]. However, event knowledge is structured knowledge with multiple participants, which has been mostly overlooked during pre-training, thus demonstrating deficiencies in tasks related to verb comprehension [62].

2.1.2 The Major Bottleneck: Cross-modal Alignment

One crucial challenge in vision-language foundation models is to align semantics across modalities at various levels of granularity, and to bridge the semantic gap between fine-grained and high-level representations. Existing pretraining models maximize the alignment across two modalities without taking into account the structure of text and images via co-attention [25, 26], label semantics [27, 29], optimal transport on a flat graph [68], etc. Image structures [4, 69] that are analogous to text-linguistic structures are proposed. There is, however, a gap between complicated linguistic structures and image structures. In this dissertation, we propose to use the text event graph structures to fill in the gap and compute a global alignment over two event graphs.

This is a new and growing area with several solutions proposed to align across modalities: (1) a hard alignment that enables granularity-aware fusion [5, 70, 71], i.e., developing sophisticated fusion and alignment mechanisms that can effectively handle semantics at different levels of granularity; (2) a soft alignment to project the text space with the vision space [50, 72, 73]. Moreover, the interdependencies among various granularities have been overlooked in a long time. A potential solution is incorporating compositional ability to transform low-level semantics into high-level visual interpretation. Such compositional capability can be potentially borrowed from linguistics to assist with the visual interpretation.

2.1.3 Knowledge-Driven Vision-Language Models

Injecting knowledge into VLMs can help connect vision and text in multiple levels of granularity. The understanding of entity knowledge (i.e., objects and object types) is the

fundamental ability for a wide variety of V+L tasks, such as image captioning [74, 75, 76, 77] and visual question answering [78, 79, 80]. They also require the capability of understanding relational knowledge (i.e., scene graphs), which can further support compositional visual question answering [81], scene graph parsing [82], etc. On top of that, event knowledge (i.e., event types, actions, activities) with event argument structures (i.e., entities involved and their semantic roles) are critical to support cognition-level visual understanding, such as visual commonsense reasoning [83], situation recognition [84, 85], action recognition [86] and human object interaction [87]. To track status changes of events and entities, procedural knowledge is induced for video question answering [88, 89], action recognition [90, 91, 92], action segmentation [93, 94], action localization [95], action prediction [96, 97, 98] and procedural planning [99]. Instead of explicitly gaining structured knowledge, the knowledge in language models can also benefit vision-language pretraining [58]. Consequently, adding knowledge into vision-language pretraining poses two key challenges: (1) obtaining knowledge at multiple levels, and (2) encoding the structure and semantics of the knowledge. In this subsection, we will review each granularity of knowledge and their usage in VLMs.

Entity-Centric Structured Knowledge for VLMs Entity-centric knowledge is the most widely used knowledge in V+L pretraining, including objects [100, 101, 102, 103] and relations (scene graphs) [104, 105, 106]. Some vision-language pretraining models learn fine-grained cross-media alignments through the internal structures of an entity regarding its bounding box features [24, 25, 26, 107] and object labels [27, 29, 108], as well as relational structures among multiple entities, such as scene graphs [34, 109, 110]. To solve the challenge of limited ontology, the frontier techniques propose to handle open-vocabulary issue through soft prompting [53, 111].

Event-Centric Structured Knowledge for VLMs Event-centric is more challenging than object-centric knowledge in terms of deep semantics and structures of multiple arguments. The key challenge of leveraging structured knowledge is structure-aware encoding, including how to introduce structures to vision modalities and how to align text and vision based upon structure. Event (activity) semantic structure extraction [84, 112, 113, 114] will be detailed in Section 2.2.2. On top of that, VLMs can learn the structural alignment of an event, including its protagonist(s), participant(s) and properties [62, 111, 115].

Procedural Knowledge for VLMs Procedural knowledge, usually represented as a script [116], is an important component to build human-level AI. It is usually defined as "knowing how to accomplish a task" and records a sequence of steps for each task. Such

knowledge could be scripted by human into an external knowledge base like wikiHow [117]. However, how to learn or leverage such knowledge from multimedia documents has been overlooked by the AI community. To inject procedural knowledge into mutlimodal models, the first step is learning procedural knowledge from instructional videos [99, 118]. It is different from traditional action anticipation [119, 120] and predictive coding [121, 122], which are not designed to learn the sequence of steps for a task. Recent efforts focus on leveraging textual knowledge base for procedural activity understanding in videos [123, 124], as well as training a neural network to learn script knowledge from multimodal data [125, 126, 127, 128].

Parametric Knowledge for VLMs Large-scale pretraining language models such as BERT [129] and GPT-3 [130] have demonstrated superior performance in capturing semantics and conducting reasoning on a wide variety of natural language tasks. Recently there is a trend to use large-scale pre-training language models to assist with understanding the semantics of vision modality. Natural language supervision are transferred to images [131] or videos [61, 132], in order to resolve commonsense understanding of visual temporal aspects [58, 59], typical visual attributes [133], visual relation parsing [134, 135], etc. Similarly, distilling knowledge from vision-language pretraining models can also facilitate language modeling [136, 137].

2.2 KNOWLEDGE EXTRACTION: EVENT EXTRACTION

Research on event extraction has been conducted independently in text and vision modalities, with different events and argument structures defined, and emphasized on distinct domains.

2.2.1 Text Event Extraction

Event extraction is proposed [138, 139]. Widely used event representation methods include event schemas [140], event knowledge graphs [141], event processes [142], event language models [143], and more recent work on event meaning representation via question-answer pairs [144, 145], event network embeddings and event time expression embeddings [146]. Text event extraction has been extensively studied for general news domain [147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158], where each event is tagged with a trigger word, and assigned to an *event type* that represents a set of synonymous verbs. Each argument is grounded to an entity, and associated with an *argument role* that the participant is playing.

There are some recent efforts that focus on jointly extracting events with entities and relations [159, 160]. DyGIE++ [161] designs a joint model to extract entities, events, and relations based on span graph propagation, while OneIE [162] further makes exploits global features to facilitate the model to capture more global interactions. Recently, indirect supervision sources have been used to improve text event extraction, include question answering and reading comprehension [163, 164, 165, 166], natural language inference [167, 168] and generation [169, 170].

Compared to vision modality, event extraction in the text modality has been researched for decades and achieved success in obtaining a situational understanding. Multimodal features have been proven to effectively improve text entity and event extraction [171].

2.2.2 Visual Event Extraction

"Events" in NLP usually refer to complex events that involve multiple entities in a large span of time (e.g. protest), while in CV [172, 173, 174] events are less complex single-entity activities (e.g. washing dishes) or actions (e.g. jumping). Visual event ontologies focus on daily life domains, such as "dogshow" and "wedding ceremony" [175]. Moreover, most efforts ignore the structure of events including arguments. There are a few methods that aim to localize the agent [176, 177, 178], or classify the recipient [179, 180, 181] of events, but neither aim to detects the complete set of arguments for an event. Another line of work similar to visual event detection is human activity detection [182] and human-object interaction [114, 180, 183, 184, 185], which focuses on the interactions between humans and objects. However, this formulation ignores the complicated structures of events, such as the argument role information and the events involving multiple participants.

The most similar task in the CV community is visual situation recognition [84, 186, 187] which aims to classify each image into one of the 500+ verbs derived from FrameNet [188] and then further classify each region as one of the 192 generic semantic roles. However, it only involves the vision modality with a focus on low-level visual details, so it lacks the ability to understand newsworthy events that requiring high-level semantic understanding. Also, the verb prediction is ambiguous since it includes various synonymous verbs, such as *protesting*, *marching*, etc, which in fact belong to the same event type with similar visual features.

Compared to static images, videos are dynamic, so video comprehension tasks not only requires the model to capture salient objects [176, 189], but also their motion [190, 191, 192, 193] and their interactions [66, 194, 195, 196, 197, 198, 199]. End-to-end models [200, 201, 202] have shown their abilities to capture certain key information for classifying action

in videos, where action is usually a single verb or a simple verb phrase [203]. Video event extraction [112] is a more complicated task than action recognition as it not only requires the model to understand the verb but also the interactions among multiple possible argument roles. It requires model to track the object state, their changes and the interaction of objects for video event extraction [17].

Due to the high variety of events and the difficulty in acquiring annotated data for training, supervised training based methods with limited pre-defined event ontologies are inapplicable in dealing with open-world events in multimedia applications. To get rid of such annotations, the key challenges in event extraction is to model the structural nature of events and their associated argument roles. The crucial aspect here is to identify the appropriate supervision that can facilitate the transfer of such structural annotations from the language modality to the vision modality.

2.2.3 Multimodal Event Extraction

Different modalities prioritize various semantic granularities. Language emphasizes high-level semantics (e.g., a rescue event), vision focuses on low-level visual details (e.g., sequence of detailed actions in the rescue event such as carrying injured person with a stretcher). Prior to Multimedia Event Extraction (M^2E^2) [4] proposed by this dissertation, event and argument extraction methods can only handle a single modality. M^2E^2 categorizes verbs into event types and requires modeling the structure of events and their arguments. It is further extended to video settings [204].

A major challenge in this line of tasks is the lack of multimedia event argument annotations, which are costly to obtain due to the annotation complexity [205]. As a result, weakly supervised frameworks [4, 206] and self-supervised training frameworks [5] are proposed to take advantage of annotated uni-modal corpora to separately learn visual and textual event extraction, and uses an image-caption dataset to align the modalities.

This task remains challenging because of two key difficulties: (1) the misalignment of semantic granularity in different modalities and (2) the complex interactions among sub-structures. The major challenge continues to be the lack of annotated training data for extracting events from multiple modalities. The only economically feasible approach to overcoming this data sparsity challenge is to develop a mechanism to effectively transfer information and knowledge from a relatively training-rich modality to another less fortunate modality.

2.3 KNOWLEDGE REASONING: EVENT SCHEME INDUCTION AND PROCEDURAL PLANNING

Understanding events requires knowledge in the form of a repository of abstracted event schemas (complex event templates). Related lines of event schema induction include narrative schema induction and script learning. Progress of events has been be researched from different angles. For example, many efforts have been devoted into modeling event narratives [143, 207] such that they can successfully predict missing events in an event process. Besides, another important event understanding angle is conceptualization [208], which aims at understanding the super-sub relations between a coarse-grained event and a fine-grained event process. In this context, the machine could also be expected to generate the event process given a goal [208], infer the goal given the process [209], and capture the recurrence of events in a process [210]. Last but not least, event coreference, which links references to the same event together, also plays a critical role in understanding events [211].

2.3.1 Script Learning

Schemas, or scripts, were originally defined by [212] as " a predetermined, stereotyped sequence of actions that defines a well-known situation". At the time of their proposal, such scripts were human curated for each scenario. Automatic schema induction had been initially thought as the discovery of event sequences governing common scenarios [213, 214]. To account for schema variations and distractions in real event instances, later work has treated schemas as probabilistic models [215, 216]. In particular, language modeling has been a popular approach to schema induction [217, 218, 219, 220, 221]. Some schemas focus on inferring the natural language steps based on the given goals [222, 223] or partially known stories [224]. Recently, [225] incorporates external commonsense knowledge bases to improve event representations, and [226] shows improvement on script induction by leveraging causal effect interventions between events.

2.3.2 Narrative Event Schema Induction

Previous work [213, 214, 217, 219, 227, 228, 229, 230, 231, 232, 233, 234] focuses on inducing *narrative schemas* as partially ordered sets of events sharing a common argument, where an event is represented as a verb with its arguments connected through typed dependencies (grammatical *subject* and *object*). It evaluates schemas via narrative cloze task [213, 217, 218, 219, 230, 231, 235] to predict the masked event in a sequence of narrative events. Event orders are inferred based on statistical events or coreferential argument co-occurrences [213,

214, 227, 229, 230], and then are further extended to include causality [236, 237], and a *temporal script graph* is proposed where events and arguments are abstracted as event types and participant types [238, 239, 240]. In this dissertation, we propose an event graph schema representation to capture more complex connections between events, and use event types instead of verbs as in previous work for more abstraction power.

2.3.3 Event Prediction

Event prediction task is designed to automatically generate a missing event (e.g., a word sequence) given a single or a sequence of prerequisite events [241, 242, 243, 244, 245], or predict a pre-condition event given the current events [246]. Previous studies predict the next events by exploiting generative seq2seq frameworks [241, 242, 244], which are incapable of capturing the temporal relations among events precisely and generalizing to diverse daily events. In contrast, we leverage the automatically discovered temporal event schema as guidance to forecast future events.

2.3.4 Graph Pattern Mining

Motif finding on heterogeneous networks [247, 248, 249, 250] discovers highly recurrent instance graph patterns, but fails in abstracting schema graphs to the type level. Previous work applies graph summarization to discover frequent subgraph patterns for heterogeneous networks [251, 252, 253, 254, 255, 256, 257, 258], focusing on the topology of the graph. However, it ignores semantic coherence among multiple patterns, and may generate disconnected subgraph patterns between events, resulting in the limited ability to show event-event connectivity through entities, and failing in generating semantically coherent patterns between events. Another line of work related to graph modeling of event schemes is graph generative models. Graph generation models can be categorized as VAE-based [259, 260, 261], flow-based [262], path-based [263, 264] and autoregressive generation models [265, 266, 267]. In this dissertation, we will introduce the first trial to use graph autoregressive model to predict events by growing the partial event graph.

2.3.5 Video-based Script Induction

Existing efforts that utilize visual information in script induction can be mainly classified into implicit script knowledge models and explicit sequential script induction models. Some previous efforts have focused on training models with implicit script knowledge that can make

step-level predictions based on textual [223], visual [124, 268, 269], or multimodal [270] input. Other models aim to produce explicit sequential graph scripts that only capture procedural relations between steps [271, 272]. Another line of works use multimodal information to generate explicit graph scripts that model only pre-conditional/dependency relationships between events [273] and sub-events [274]. Ours is the first work to generate explicit non-sequential graph scripts that capture rich procedural, optional, and interchangeable relations through multimodal learning.

2.3.6 Instructional Procedure Planning

Introduced by [99], the procedure planning task aims at predicting the intermediate steps (actions) given a start visual observation and a goal visual observation. The key challenge of this task lies in its unstructured, highly diverse observations which are unsuitable for directly planning over. To tackle this challenge, most previous approaches [99, 275, 276, 277] attempt to learn a latent space from visual observations by a supervised imitation learning objective over both the actions and the intermediate visual observations. More recently, P3IV[278] observes that actions can be treated as both discrete labels and natural language. By using a pretrained vision-language model to encode the actions as text, P3IV achieves higher planning success rate using only action-level supervision. P3IV can be seen as an attempt to map the action text into visual space to provide more stable supervision. In comparison, our model maps visual observations into text space.

2.3.7 Pre-trained Language Models for Planning

Recent work has shown the potential of language models for text-based planning tasks. Language models pre-trained on a large internet-scale corpus encodes rich semantic knowledge about the world and are equipped with strong low-shot reasoning abilities. In the effort of connecting language models with embodied AI, pioneering work on text-based planning [279, 280, 281] shows that learning to solve tasks using abstract language as a starting point can be more effective and generalizable than learning directly from embodied environments. More recently, [282, 283, 284, 285] further show that using large language models as out-of-the-box planners brings significant benefits to a wide range of embodied tasks, such as navigation and instruction following.

2.4 KNOWLEDGE-DRIVEN GENERATION

2.4.1 Event Summarization

Over the past year, we saw many examples of such events, including COVID-19, the vaccine roll-out, the Black Lives Matter movement and the US presidential election. In this tutorial, we will present methods for tracking such events over time and generating summaries that provide updates as an event unfolds. The task of identifying and tracking events was first introduced in the Topic Detection and Tracking challenge [286]. Recent work has explored new methods for tracking and visualizing such events over time (e.g., [287, 288, 289, 290]), in some cases generating summaries that contain information on what is new (e.g., [291, 292]) and in other cases, exploring timeline summarization, ordering events and generating summaries that are placed along a timeline (e.g., [293, 294, 295, 296]) We will also consider how these are related to summarization of an event that takes place within a single day, a problem that falls within the category of multi-document summarization (e.g., [297, 298]), as typically there may be many articles covering the same event. By using multiple articles as input, a summarizer can present different perspectives on the same event as well as identify salient information that is highlighted many in different ways across the set of input articles.

2.4.2 Timeline Summarization

Due to the lack of training data, timeline summarization focuses on extractive methods with heuristics [294, 296, 299, 300, 301, 302, 303, 304, 305], with a few abstractive methods [295, 306, 307] that require a few gold summaries to work. They both fail to capture the rich event structures and ignore the temporal orders between events. We are the first to use optimal transport on summarization task to select semantic relevant, structurally salient and temporally coherent events.

Graph-based Multi-Document Summarization methods either extractive [297, 308, 309, 310, 311], or abstractive [312, 313, 314, 315]. They are closely related to timeline summarization but cannot be directly applied, due to the lack of temporal dimensions in generation.

2.4.3 Graph Representation of Documents.

In general NLP research, people have built various text graphs by augmenting original text sequences with different hidden structural information, such as entity-centric graphs for

efficient joint-encoding of large corpora [316, 317, 318, 319, 320, 321]. Event graphs from a single document have been built for event schema induction [243, 322], event coreference resolution [19, 323], etc. However, they ignore relations between event arguments, or only use hierarchical or temporal relations to connect events. Also, cross-document entity coreference and event coreference resolution are critical for large corpora understanding, while previous work focuses on a single document. Our approach is unique in building event-centric graphs across documents, with rich argument and temporal information.

CHAPTER 3: COMPLEXITY: MULTIMODAL SEMANTIC MODELING

The first step of event-centric understanding is discovering events from multimodal data. Multimedia presentation enables us to obtain a comprehensive and holistic understanding by fusing information from multiple modalities. For example, by randomly sampling 100 multimodal news articles from the Voice of America (VOA) [324], we find that 33% of images in the articles contain visual objects that serve as event arguments and are not mentioned in the text [4]. In this chapter, we study the joint extraction of event structures from text and vision modalities, and close the gap by aligning structures from two modalities.

The extraction of factual knowledge about events has been independently researched in text and vision. Regarding the language modality, Natural Language Processing (NLP) has experienced great successes in text-only event extraction. Recent research in vision activities or situations can be regarded as event extraction in Computer Vision (CV), but with major differences in task definition, data domain, methodology, and focus. However, a comprehensive understanding of events requires computers to perform joint comprehension across multiple modalities, such as text, images, and videos. Therefore, we propose a new task [4] to learn such an ability, and propose both supervised [4] and self-supervised [111] methods to tackle multimodal event extraction.

3.1 A NEW TASK OF MULTIMODAL EVENT EXTRACTION (M^2E^2)

Traditional event extraction methods target a single modality, such as text [161], images [84] or videos [66, 325, 326]. However, the practice of contemporary journalism [1] distributes news content via multimedia, as people often can better understand complex events through both text and vision.

In this dissertation, we introduce a new task, **M**ulti**M**edia **E**vent **E**xtraction (M^2E^2), which aims to extract events and their arguments from multimodal documents (*e.g.,* news articles consisting of text and images). In this task, events are coreferential across modalities. It develops a common abstraction across modalities, i.e., an extracted event can come from text, image, or both, and the event arguments can also come from any modality. Taking Figure 3.1 as an example, the AGENT and PERSON arguments of the MOVEMENT.TRANSPORT event are mentioned in the text, but the VEHICLE argument (the visual object *truck*) is only visible in the image.

Formally, each input document consists of a set of images $\mathcal{M} = \{m_1, m_2, \dots\}$ and a set of sentences $\mathcal{S} = \{s_1, s_2, \dots\}$. Each sentence s can be represented as a sequence of tokens

$s = (w_1, w_2, \dots)$, where w_i is a token from the document vocabulary $\mathcal{W}$. The input also includes a set of entities $\mathcal{T} = \{t_1, t_2, \dots\}$ extracted from the document text. An entity is an individually unique object in the real world, such as a person, an organization, a facility, a location, a geopolitical entity, a weapon, or a vehicle. The objective of M^{2}E^{2}is two-fold:

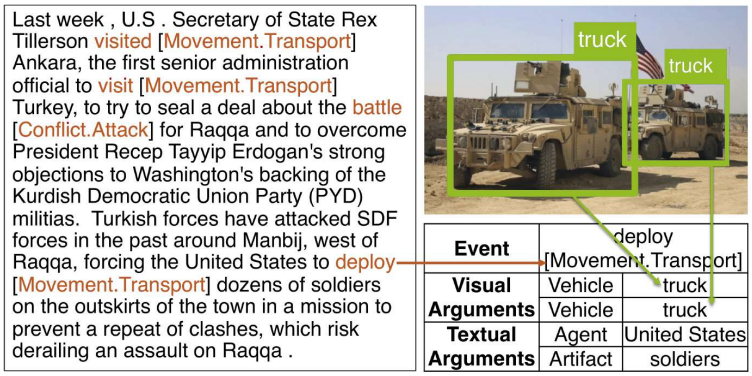


Figure 3.1: An example of Multimodal Event Extraction.

(1) Event Extraction: Given a multimodal document, extract a set of event mentions, where each event mention e has a type y_e and is grounded on a text trigger word w or an image m or both,

$$e = (y_e, \{w, m\}). \tag{3.1}$$

Note that for an event, w and m can both exist, which means the visual event mention and the textual event mention refer to the same event. For example in Figure 3.1, *deploy* indicates the same MOVEMENT.TRANSPORT event as the image. y_0 is *Movement.Transport*, G_0^T is $\{deploy\}$, and G_0^M includes the image. We consider the event e as **text-only** event if it only has textual mention w, and as **image-only** event if it only contains visual mention m, and as **multimodal** event if both w and m exist.

(2) Argument Extraction: The second task is to extract a set of arguments of event mention e. Each argument a has an argument role type y_a, and is grounded on a text entity t or an image object o (represented as a bounding box), or both,

$$a = (y_a, \{t, o\}). \tag{3.2}$$

The arguments of visual and textual event mentions are merged if they refer to the same real-world event, as shown in Figure 3.1.

Event Type	**Definition**
Movement.Transport	it occurs when an *artifact* or a *person* is moved from one place to another
Conflict.Attack	a violent physical act causing harm or damage
Conflict.Demonstrate	it occurs when a large number of people come together in a public area to protest or demand
Justice.ArrestJail	it occurs when the movement of a person is constrained by a state actor
Contact.PhoneWrite	it occurs when two or more people directly engage in discussion but not face-to-face
Contact.Meet	it occurs when two or more people interact with one another face-to-face at a single location
Life.Die	it occurs when the life of a person entity ends
Transaction.TransferMoney	it refers to the giving, receiving, borrowing, or lending money

Table 3.1: Event type definition in M^2E^2 dataset.

3.2 DATASET CONSTRUCTION FOR MULTIMODAL EVENT LEARNING

3.2.1 $\mathbf{M^2E^2}$: A New Benchmark for Multimodal Event Evaluation

We define multimedia newsworthy event types by exhaustively mapping between the event ontology in NLP community for the news domain (ACE[1]) and the event ontology in CV community for general domain (imSitu [84]). They cover the largest event training resources in each community. Table 3.1 shows the selected complete intersection, which contains 8 ACE types (i.e., 24% of all ACE types), mapped to 98 imSitu types (i.e., 20% of all imSitu types). We expand the ACE event role set by adding visual arguments from imSitu, such as *instrument*, bolded in Table 3.2. Here, numbers in parentheses represent the counts of textual and visual events/arguments. This set encompasses 52% ACE events in a news corpus, which indicates that the selected eight types are salient in the news domain. While the dataset is feasibly extensible to any types that can be represented in both modalities, the current set of eight event types is selected by exhaustively mapping the ACE and SR ontologies. We reuse these existing ontologies because they enable us to train event and argument classifiers for both modalities without requiring joint multimedia event annotation as training data.

[1] `https://catalog.ldc.upenn.edu/ldc2006T06`

We collect 108,693 multimedia news articles from the Voice of America (VOA) website [2] 2006-2017, covering a wide range of newsworthy topics such as military, economy and health. We select 245 documents as the annotation set based on three criteria: (1) Informativeness: articles with more event mentions; (2) Illustration: articles with more images (> 4); (3) Diversity: articles that balance the event type distribution regardless of true frequency. For the first and third criteria, we use the baseline text-only event extraction model [327] to estimate the number of event mentions per event type in each articles.

Event Type	**Argument Role**
Movement.Transport (223|53)	Agent (46|64), Artifact (179|103), Vehicle (24|51), Destination (120|0), Origin (66|0)
Conflict.Attack (326|27)	Attacker (192|12), Target (207|19), Instrument (37|15), Place (121|0)
Conflict.Demonstrate (151|69)	Entity (102|184), **Police** (3|26), **Instrument** (0|118), Place (86|25)
Justice.ArrestJail (160|56)	Agent (64|119), Person (147|99), **Instrument** (0|11), Place (43|0)
Contact.PhoneWrite (33|37)	Entity (33|46), **Instrument** (0|43), Place (8|0)
Contact.Meet (127|79)	Participant (119|321), Place (68|0)
Life.Die (244|64)	Agent (39|0), Instrument (4|2), Victim (165|155), Place (54|0)
Transaction.TransferMoney (33|6)	Giver (19|3), Recipient (19|5), Money (0|8)

Table 3.2: Event types and argument roles in M^2E^2.

The data statistics are shown in Table 3.3. Among all of these events, 192 textual event mentions and 203 visual event mentions can be aligned as 309 cross-media event mention pairs. The dataset can be divided into 1,105 text-only event mentions, 188 image-only event mentions, and 395 multimedia event mentions.

We follow the ACE event annotation guidelines[328] for textual event and argument annotation, and design an annotation guideline [3] for multimedia events annotation.

We annotate event type and argument roles for textual and visual events. The annotation process involves tasks in Table 3.4. After completing text-independent and image-

[2] `https://www.voanews.com/`

[3] `http://blender.cs.illinois.edu/software/m2e2/ACL2020_M2E2_annotation.pdf`

Source		Event Mention		Argument Role	
sentence	image	textual	visual	textual	visual
6,167	1,014	1,297	391	1,965	1,429

Table 3.3: Data statistics of M^2E^2.

Modality	Type	Task
Text	Event Type	Classification (Event Type)
		Localization (Trigger)
	Argument	Classification (Argument)
		Localization (Entity)
Image	Event Type	Classification (Event Type)
	Argument	Classification (Argument)
		Localization (Union)
		Localization (Instance)
Cross	Coreference	Classification (Relation)

Table 3.4: M^2E^2 annotation tasks.

independent annotations, expert annotators are asked to perform adjudication. We do **not** tag all events, but only a particular subset from ACE ontology, as in Table 3.1.

Textual event annotation includes event type annotation and argument annotation. we assign an event type to each event trigger (the words or phrases that most clearly express event occurrences), and an argument role to each participant (entity, time or value expression). Here we focus on intra-sentence event extraction and do not consider cross-sentence or cross-document situations.

Visual event annotation includes event type annotation and argument annotation. We assign an event type to each image if the image contains pre-defined types of events, and assign argument roles to corresponding bounding boxes. The event type annotation does not locate a specific region in the image, but use the whole image as justification.

After annotating events and arguments separately in each modality, we ask annotators to find image-sentence pairs that correspond to the same event instance, i.e., the same event type happening in the same location and time.

This guideline focuses on how to annotate events and argument roles in images. For more details in text event annotation, please refer to the *ACE English Annotation Guidelines for*

Events (Version 5.4.3 2005.07.01)[4].

One challenge is that we often need the surrounding texts to disambiguate event argument roles in images. Therefore, we require each annotator to label each image by checking the caption as reference context.

Figure 3.2: Examples of union and instance bounding boxes.

Another unique challenge in multimedia event annotation is to localize visual arguments in complex scenarios, where images include a crowd of people or a group of object. It is hard to delineate each of them using a bounding box. To solve this problem, we define two types of bounding boxes:

(1) *union bounding box*: for each role, we annotate the smallest bounding box covering all constituents;

(2) *instance bounding box*: for each role, we annotate a set of bounding boxes, where each box is the smallest region that covers an individual participant (e.g., one person in the crowd), following the VOC2011 Annotation Guidelines[5].

An example is shown in Figure 3.2. Annotation is done via two independent passes by 8 NLP and CV researchers, and 2 expert annotators perform adjudication. After annotating events and arguments separately for each modality, we ask annotators to find image-sentence pairs that correspond to the same event instance.

3.2.2 VOANews: A New Event-Rich Dataset for Pretraining

We collect 106,875 image-captions that are rich in events from news websites [324]. It provides a new challenging image-retrieval benchmark, where each sentence may contain multiple events with a complicated linguistic structure. The average caption length is 28.3 tokens, compared to 13.4 for Flickr30k and 11.3 for MSCOCO. The data statistics are shown in Table 3.5. Structural event knowledge extracted automatically following Section 3.4.1.

[4] `https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf`

[5] `http://host.robots.ox.ac.uk/pascal/VOC/voc2011/guidelines.html`

The type distribution is shown in Figure 3.3, indicating the event types in our collected dataset are generally visually detectable.

Dataset	Split	#image	#event	#arg	#ent
VOANews	**Train**	76,256	84,120	148,262	573,016
	Test	18,310	21,211	39,375	87,671
	No-event	12,309	-	-	-

Table 3.5: Data statistics of VOANews.

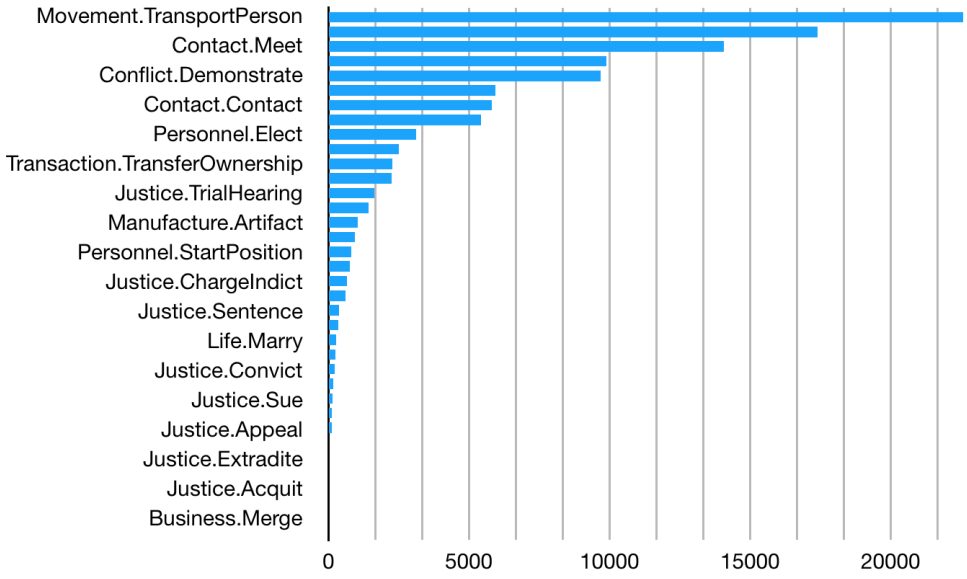


Figure 3.3: Event distribution in VOANews.

3.3 SUPERVISED MULTIMODAL EVENT EXTRACTION

As shown in Figure 3.4, the training phase contains three tasks: text event extraction, visual situation recognition, and cross-media alignment. We learn a cross-media shared encoder, a shared event classifier, and a shared argument classifier. In the testing phase, given a multimodal news article, we encode the sentences and images into the structured common space, and jointly extract textual and visual events and arguments, followed by cross-modal coreference resolution.

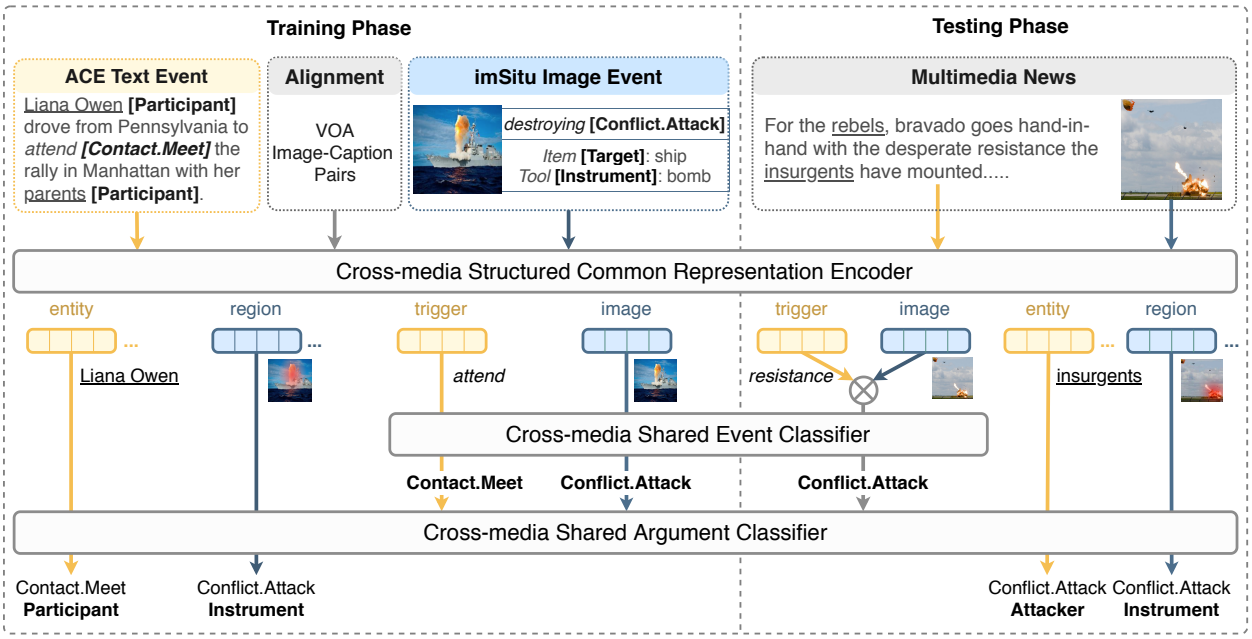


Figure 3.4: Approach overview of supervised multimodal event extraction in training and testing phases.

3.3.1 Text Event Extraction

We choose Abstract Meaning Representation (AMR) [329] to represent text, and apply a Graph Convolutional Network (GCN) [330] to encode the graph contextual information following [156]. We classify each word w into event types y_e[6] and classify each entity t into argument role y_a:

$$P(y_e|w) = \frac{\exp\left(\boldsymbol{W}_e \boldsymbol{w}^{\mathbb{C}} + \boldsymbol{b}_e\right)}{\sum_{e'} \exp\left(\boldsymbol{W}_{e'} \boldsymbol{w}^{\mathbb{C}} + \boldsymbol{b}_{e'}\right)}, P(y_a|t) = \frac{\exp(\boldsymbol{W}_a[\boldsymbol{t}^{\mathbb{C}}; \boldsymbol{w}^{\mathbb{C}}] + \boldsymbol{b}_a)}{\sum_{a'} \exp(\boldsymbol{W}_{a'}[\boldsymbol{t}^{\mathbb{C}}; \boldsymbol{w}^{\mathbb{C}}] + \boldsymbol{b}_{a'})}. \quad (3.3)$$

3.3.2 Image Event Extraction

To obtain image structures similar to AMR graphs, and inspired by *situation recognition* [84], we represent each image with a *situation graph*, that is a star-shaped graph as shown in Figure 3.5, where the central node is labeled as a verb v (e.g., *destroying*), and the neighbor nodes are arguments labeled as $\{(n, r)\}$, where n is a noun (e.g., *ship*) derived from WordNet synsets [331] to indicate the entity type, and r indicates the role (e.g., *item*) played by the entity in the event, based on FrameNet [332].

[6]We use BIO tag schema to decide trigger word boundary, i.e., adding prefix *B-* to the type label to mark the beginning of a trigger, *I-* for inside, and *O* for none.

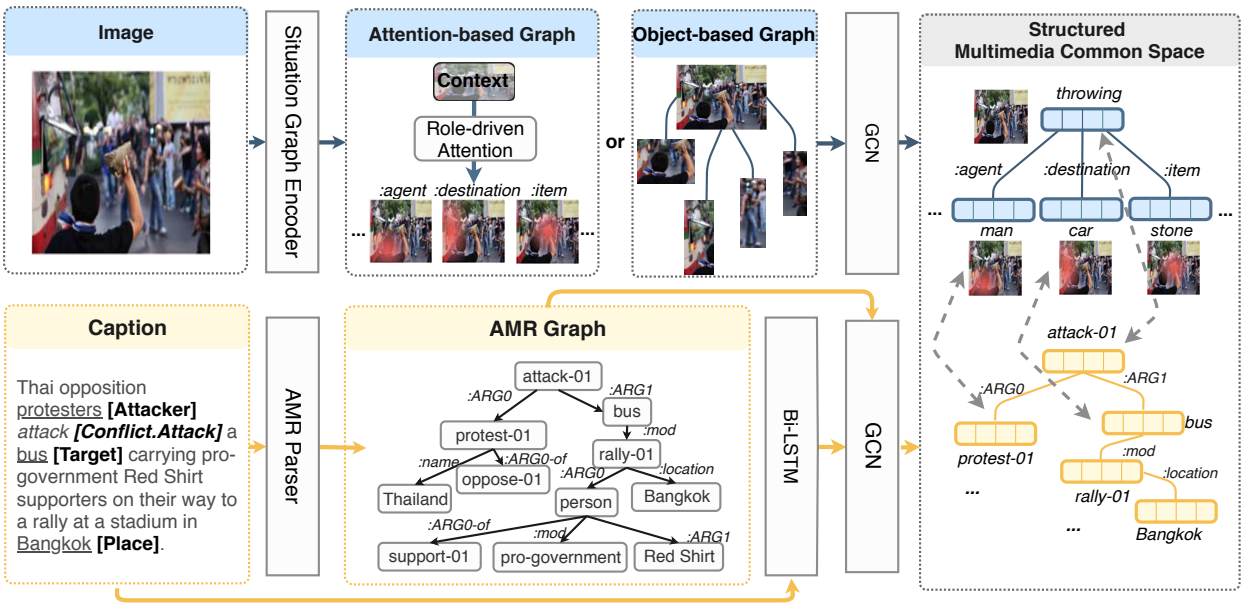


Figure 3.5: Multimodal structured common space construction.

We use pre-train situation graph parser on the imSitu dataset [84]. Then we apply a GCN to obtain the structured embedding of each node in the common space. This yields $\boldsymbol{m}^{\mathbb{C}}$ and $\boldsymbol{o}_i^{\mathbb{C}}$. We use the same classifiers as defined in Equation 3.3 to classify each visual event and argument using the common space embedding:

$$P(y_e|m) = \frac{\exp(\boldsymbol{W}_e \boldsymbol{m}^{\mathbb{C}} + \boldsymbol{b}_e)}{\sum_{e'} \exp(\boldsymbol{W}_{e'} \boldsymbol{m}^{\mathbb{C}} + \boldsymbol{b}_{e'})}, P(y_a|o) = \frac{\exp(\boldsymbol{W}_a [\boldsymbol{o}^{\mathbb{C}}; \boldsymbol{m}^{\mathbb{C}}] + \boldsymbol{b}_a)}{\sum_{a'} \exp(\boldsymbol{W}_{a'} [\boldsymbol{o}^{\mathbb{C}}; \boldsymbol{m}^{\mathbb{C}}] + \boldsymbol{b}_{a'})}. \tag{3.4}$$

3.3.3 Cross-Media Joint Training.

Since there is no ground truth alignment between the image nodes and caption nodes, we use image and caption pairs for weakly supervised training, to learn a soft alignment from each words to image objects and vice versa.

$$\alpha_{ij} = \frac{\exp\left(\boldsymbol{w}_i^{\mathbb{C}} \boldsymbol{o}_j^{\mathbb{C}}\right)}{\sum_{j'} \exp\left(\boldsymbol{w}_i^{\mathbb{C}} \boldsymbol{o}_{j'}^{\mathbb{C}}\right)}, \quad \beta_{ji} = \frac{\exp\left(\boldsymbol{w}_i^{\mathbb{C}} \boldsymbol{o}_j^{\mathbb{C}}\right)}{\sum_{i'} \exp\left(\boldsymbol{w}_{i'}^{\mathbb{C}} \boldsymbol{o}_j^{\mathbb{C}}\right)}, \tag{3.5}$$

where w_i indicates the i^{th} word in caption sentence s and o_j represents the j^{th} object of image m. Then, we compute a weighted average of softly aligned nodes for each node in other modality,

$$\boldsymbol{w}_i' = \sum_j \alpha_{ij} \boldsymbol{o}_j^{\mathbb{C}}, \ \boldsymbol{o}_j' = \sum_i \beta_{ji} \boldsymbol{w}_i^{\mathbb{C}}. \tag{3.6}$$

We define the alignment cost of the image-caption pair as the Euclidean distance between each node to its aligned representation,

$$\langle s, m \rangle = \sum_i ||\boldsymbol{w}_i - \boldsymbol{w}'_i||_2^2 + \sum_j ||\boldsymbol{o}_j - \boldsymbol{o}'_j||_2^2 \tag{3.7}$$

We use a triplet loss to pull relevant image-caption pairs close while pushing irrelevant ones apart:

$$\mathcal{L}_c = \max(0, 1 + \langle s, m \rangle - \langle s, m^- \rangle), \tag{3.8}$$

where m^- is a randomly sampled negative image that does not match s. Note that in order to learn the alignment between the image and the trigger word, we treat the image as a special object when learning cross-media alignment.

The common space enables the event and argument classifiers to share weights across modalities, and be trained jointly on the ACE and imSitu datasets, by minimizing three tasks jointly.

3.4 ZERO-SHOT MULTIMODAL EVENT EXTRACTION

Inspired by the great success of zero-shot object detection [333] using vision-language pretraining models [26, 29, 30], we leverage the power of vision-language pretraining models to tackle Zero-shot Multimodal Event Extraction. However, existing vision-language pretraining models [25, 26, 27, 29, 30, 31] focus on the understanding of images or entities, ignoring the event semantics and structures. As a result, apparent failures happen in the circumstances requiring verb comprehension [62]. Thus, we focus on integrating event structural knowledge into vision-language pretraining.

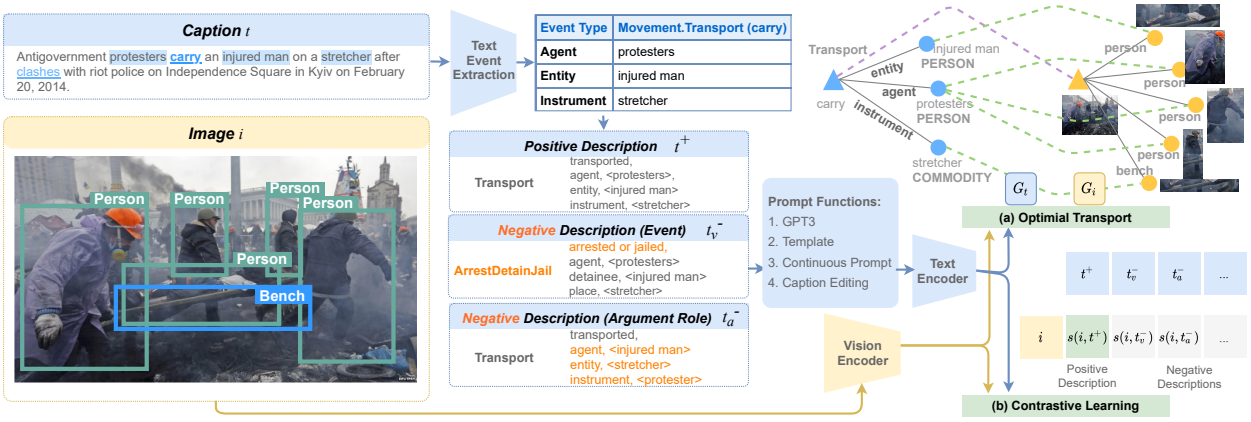


Figure 3.6: Architecture of CLIP-Event.

Previous work primarily represents visual events as verbs with subjects and objects [114, 180, 182, 183, 184, 185]. However, events contain structural knowledge, with each event being assigned to an *event type* that represents a set of synonymous verbs. Each argument is grounded to text or images, and associated with an *argument role* that the participant is playing. As shown in Figure 3.6, the *carry* event is typed as TRANSPORT, with *protesters* as AGENT, *injured man* as ENTITY and *stretcher* as INSTRUMENT.

Our goal is to incorporate event structured knowledge into vision-language pretraining. In the following we will address two primary questions regarding the model design: (1) How can the structural event knowledge be acquired? (2) How can the semantics and structures of events be encoded?

3.4.1 Event Structural Knowledge Extraction

Text and Visual Knowledge Extraction We use a state-of-the-art text information extraction system [162] to obtain text event structures and use it as the supervision for the given image, as shown in Figure 3.7. In detail, we run the dockerized version [7] to extract events of 187 types[7], covering a wide range of newsworthy events. For images, we apply Faster R-CNN [101] trained on Open Images [334] to detect objects.

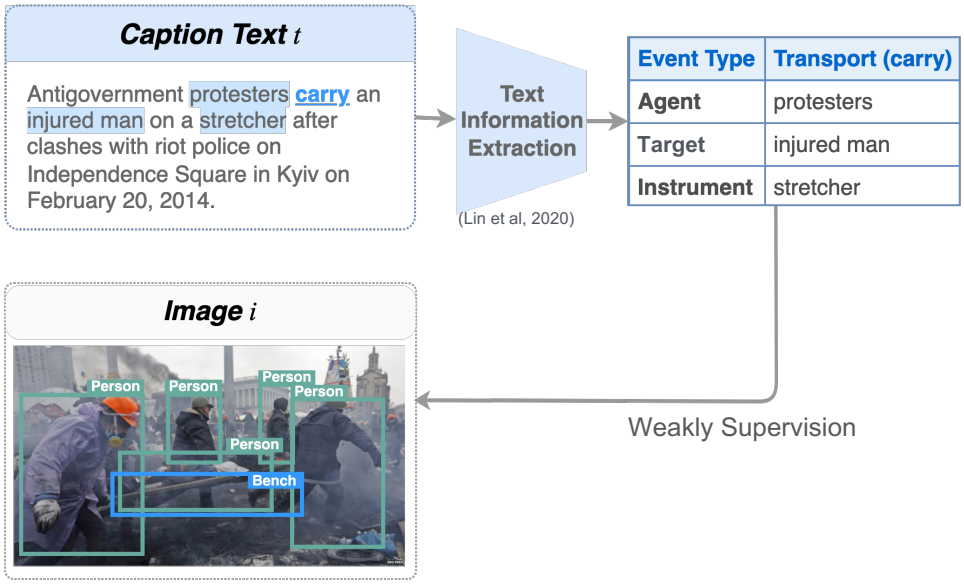


Figure 3.7: Transfer text event knowledge to images.

[7]The system uses DARPA AIDA ontology, which is the most fine-grained text event ontology, available at `https://github.com/NextCenturyCorporation/AIDA-Interchange-Format/blob/master/java/src/main/resources/com/ncc/aif/ontologies/LDCOntologyM36`.

Primary Event Detection When there are multiple events in the caption, the image typically depicts the primary event of the caption. We detect the primary event as the event that is closer to the root of dependency parsing tree [335], and has a larger number of arguments, higher event type frequency, and higher similarity between trigger word and the image using the pretrained CLIP model [30]. We rank events according to these criteria, and perform majority voting. For example, in Figure 3.6, there are two events *carry* and *clashes* in the caption. We select *carry* as the primary event since it is the root of the dependency tree, and it has three arguments, as well as higher similarity with the image.

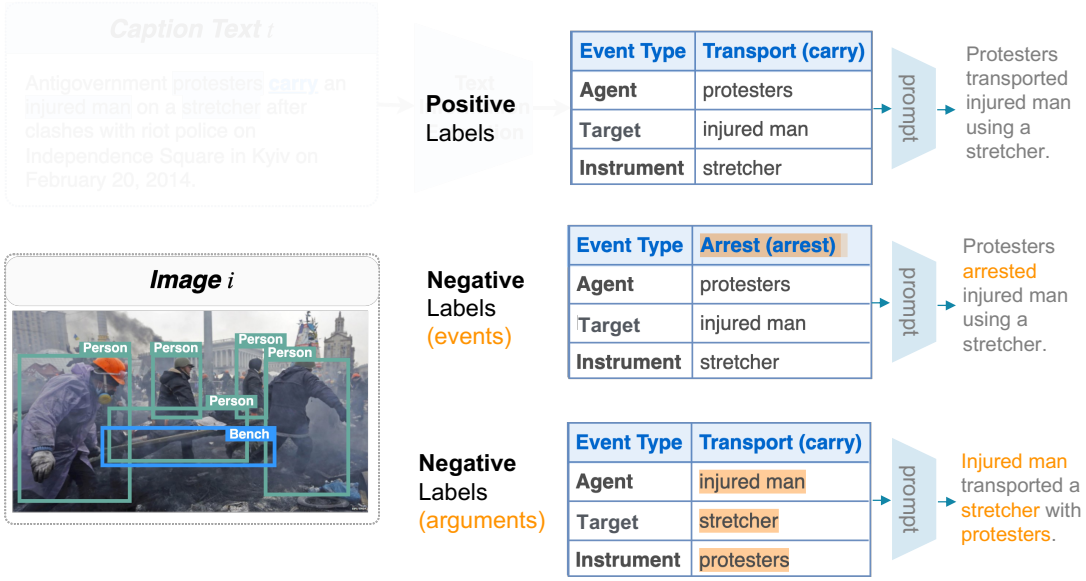


Figure 3.8: Hard negative construction via manipulating text event structures.

3.4.2 Event Structure Driven Negative Sampling

To force the Text and Vision Encoders to learn robust features about event types and argument roles, we design the following strategies to generate challenging negatives.

Negative Event Sampling We compute the confusion matrix for the event type classifier of the state-of-the-art vision-language pretraining model CLIP [30] on the pretraining image-caption dataset. The classifier is based on the similarity scores between the event type labels $\phi_v \in \Phi_V$ (such as TRANSPORT) and the input image i, and select the top one as the predicted event type $\phi_v^\star$.

$$\phi_v^\star = \arg\max_{\phi_v \in \Phi_V} \boldsymbol{\phi_v}^T \cdot \boldsymbol{i}, \tag{3.9}$$

Prompt		Example descriptions of Figure 3.6 with arrest as negative event
Single Template	Template	⟨arg1⟩ **transported** ⟨arg2⟩ in ⟨arg3⟩ instrument from ⟨arg4⟩ place to ⟨arg5⟩ place.
	Positive	*Protesters* **transported** *an injured man* in *a stretcher* instrument.
	Negative-Evt	*Protesters* **arrested** *an injured man* in *a stretcher* place.
	Negative-Arg	*An injured man* **transported** *a stretcher* in *protesters* instrument.
Composed Template	Template	The image is about **Transport**. The AGENT is ⟨arg1⟩. The ENTITY is ⟨arg2⟩. The INSTRUMENT in ⟨arg3⟩. The ORIGIN is ⟨arg4⟩. The DESTINATION is ⟨arg5⟩.
	Positive	The image is about **Transport**. The AGENT is *protesters*. The ENTITY is *an injured man*. The INSTRUMENT is *a stretcher*.
	Negative-Evt	The image is about **Arrest**. The AGENT is *protesters*. The DETAINEE is *an injured man*. The PLACE is *a stretcher*.
	Negative-Arg	The image is about **Transport**. The AGENT is *an injured man*. The ENTITY is *a stretcher*. The INSTRUMENT is *protesters*.
Continuous Prompt	Template	$[X_0]$ **Transport** $[X_1]$ AGENT $[X_2]$ ⟨arg1⟩ $[X_3]$ ENTITY $[X_2]$ ⟨arg2⟩ $[X_3]$ INSTRUMENT $[X_2]$ ⟨arg3⟩ $[X_3]$ ORIGIN $[X_2]$ ⟨arg4⟩ $[X_3]$ DESTINATION $[X_2]$ ⟨arg5⟩ $[X_3]$
	Positive	$[X_0]$ **Transport** $[X_1]$ AGENT $[X_2]$ *protesters* $[X_3]$ ENTITY $[X_2]$ *an injured man* $[X_3]$ INSTRUMENT $[X_2]$ *a stretcher* $[X_3]$
	Negative-Evt	$[X_0]$ **Arrest** $[X_1]$ AGENT $[X_2]$ *protesters* $[X_3]$ DETAINEE $[X_2]$ *an injured man* $[X_3]$ PLACE $[X_2]$ *a stretcher* $[X_3]$
	Negative-Arg	$[X_0]$ **Transport** $[X_1]$ AGENT $[X_2]$ *an injured man* $[X_3]$ ENTITY $[X_2]$ *a stretcher* $[X_3]$ INSTRUMENT $[X_2]$ *protesters* $[X_3]$
Caption Editing	Positive	*Antigovernment protesters* **carry** *an injured man* on *a stretcher* after clashes with riot police on Independence Square in ...
	Negative-Evt	*Antigovernment protesters* **arrest** *an injured man* on *a stretcher* after clashes with riot police on Independence Square in ...
	Negative-Arg	*An injured man* **carry** *a stretcher* on *antigovernment protesters* after clashes with riot police on Independence Square in ...
GPT-3	Positive	*Protesters* **transported** *an injured man* with *a stretcher*.
	Negative-Evt	*Protesters* **arrested** *an injured man* with *a stretcher*.
	Negative-Arg	*An injured man* **transported** *a stretcher* and *protesters*.

Table 3.6: The automatically generated positive and negative descriptions in CLIP-Event.

where the bold symbols stand for the representations from the Text and Vision Encoders in Figure 3.6, and we follow CLIP to use Text and Vision Transformers. The confusion matrix is computed by comparing the predicted event type with the type of the primary

event for the image. As a result, the negative event types are the challenging cases in image event typing, i.e., the event types whose visual features are ambiguous with the primary event type. For example, in Figure 3.6, ARREST is sampled as a negative event type, since its visual features are similar to TRANSPORT.

Negative Argument Sampling For argument roles, since each event by definition has multiple arguments, we manipulate the order of arguments by performing a right-rotation of the argument role sequence. In detail, we first order existing argument roles following the ontology definition, such as "AGENT, ENTITY, INSTRUMENT" in Figure 3.6. After that, we right rotate the argument role sequence by one step, resulting in "INSTRUMENT, AGENT, ENTITY". As a result, each argument is re-assigned to a manipulated role, e.g., *injured man*, the second argument, is manipulated from ENTITY to AGENT. If there is only one argument for the event, we sample a negative role according to the argument confusion matrix of the text argument extraction system [162].

Description Generation To encode the positive and negative event structures using the Text Encoder, we design multiple prompt functions, as shown in Table 3.6:

(1) **Single Template-based Prompt** encodes all arguments in one sentence.

(2) **Composed Template-based Prompt** uses a short sentence to each argument.

(3) **Continuous Prompt** employs learnable prepended tokens `[X`$_i$`]`.

(4) **Caption Editing** has minimum information loss by only altering event trigger word or switching arguments.

(5) **GPT-3 based Prompt** generates a semantically coherent natural language description conditioned on the event structure. We employ GPT-3 [336] and use five manual event description examples as few-shot prompts [336] to control the generation. The input to GPT-3 is the concatenation of the example events (`[ex_v]`) with arguments (`[ex_a]`), the example descriptions (`[ex_desp]`), and the target events (`[input_v]`) with arguments (`[input_a]`). The output of GPT-3 is the target description (`[output_desp]`). The description is more natural compared to template-based methods.

3.4.3 Event Graph Alignment via Optimal Transport

Each event and its arguments can be organized as a graph, as shown in Figure 3.6, where the central node is the event node (triangle nodes), and it's connected to entities (circle nodes) via argument roles. Encoding event graph structures enables the model to capture

the interactions between events and arguments. For example, the *injured man* should be aligned with the ENTITY being transported, rather than the AGENT.

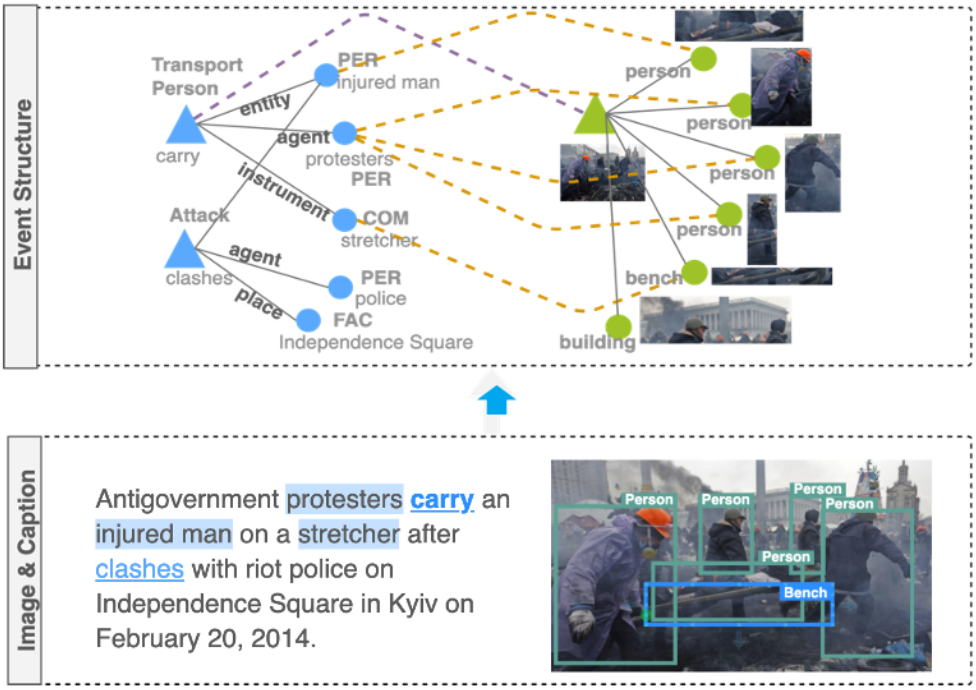


Figure 3.9: Graph alignment between image event graph and text event graph.

Image-level Alignment We compute cosine similarity $s(t, i)$ and distance $d(t, i)$ between the text t and image i:

$$s(t, i) = \cos(\boldsymbol{t}, \boldsymbol{i}), d(t, i) = c(\boldsymbol{t}, \boldsymbol{i}), \tag{3.10}$$

where $c(\cdot, \cdot) = 1 - \cos(\cdot, \cdot)$ is the cosine distance function, and $\boldsymbol{t}$ is obtained from the Text Transformer and $\boldsymbol{i}$ is obtained from the Vision Transformer.

Entity-level Alignment The cosine distance between text entity e and image object o considers both the mention similarity and type similarity.

$$d(e, o) = c(\boldsymbol{t_e}, \boldsymbol{i_o}) + c(\boldsymbol{\phi_e}, \boldsymbol{\phi_o}), \tag{3.11}$$

where t_e is the text mention of entity e, and $\boldsymbol{t_e}$ is its embedding contextualized on the sentence. We encode the sentence using the Text Transformer following [30], and apply average pooling over the tokens in the entity mention t_e. Similarly, i_o is the bounding box of object o and $\boldsymbol{i_o}$ is its embedding contextualized on the image, based on the average pooling

over the Vision Transformer representations of the patches covered in the bounding box. ϕ_e and ϕ_o are the type representations encoded by the Text Transformer. For example, $\phi_e = \textsc{person}$ for $e =$ *injured man* and $\phi_o = \textsc{person}$ for $o =$  . Therefore, the distance between the aforementioned entity and object is:

$$d(e, o) = c(\textit{injured man}, \text{  }) + c(\textsc{Person}, \textsc{Person}), \tag{3.12}$$

Event-level Alignment To obtain a global alignment score based on the structures of two graphs, we use the optimal transport [337] to get the minimal distance $d(G_t, G_i)$ between text event graph G_t and image event graph G_i,

$$d(G_t, G_i) = \min_{\boldsymbol{T}} \boldsymbol{T} \odot \boldsymbol{C}, \tag{3.13}$$

where $\odot$ represents the Hadamard product. $\boldsymbol{T} \in \mathbb{R}_+^{n \times m}$ denotes the transport plan, learned to optimize a *soft* node alignment between two graphs. n and m are the numbers of nodes in G_t and G_i, respectively. Namely, each node in text graph G_t can be transferred to multiple nodes in image graph G_i with different weights.

C is the cost matrix. We define cost between event nodes, and between argument nodes. For event nodes, the cost is the cosine distance between the image i and trigger word v,

$$C(v, i) = c(\boldsymbol{t_v}, \boldsymbol{i}) + c(\boldsymbol{\phi_v}, \boldsymbol{i}). \tag{3.14}$$

For example, in Figure 3.6, $v =$ *carry* and $\phi_v = \textsc{Transport}$,

$$C(v, i) = c(\textit{carry}, \text{  }) + c(\textsc{Transport}, \text{  }). \tag{3.15}$$

The representation $\boldsymbol{t}_v$ is also from the Text Transformer, contextualized on the text sentence.

The cost between each argument $\langle a, e \rangle$ and each bounding box o is based on the similarity of object o with both argument role a and text entity e.

$$\begin{aligned} C(\langle a, e \rangle, o) &= d(a, o) + d(e, o) \\ &= c(\boldsymbol{t_a}, \boldsymbol{i_o}) + c(\boldsymbol{t_e}, \boldsymbol{i_o}) + c(\boldsymbol{\phi_e}, \boldsymbol{\phi_o}), \end{aligned} \tag{3.16}$$

where t_a is the argument description. For example, for the argument role $a = \textsc{Entity}$ of entity $e =$ *injured man*,

$$\begin{aligned} C(\langle a, e \rangle, o) = c(&\textsc{Entity } \textit{of } \textsc{Transport}, \text{  }) \\ &+ c(\textit{injured man}, \text{  }) + c(\textsc{Person}, \textsc{Person}). \end{aligned} \tag{3.17}$$

The optimal $\boldsymbol{T} \in \mathbb{R}_+^{n \times m}$ that solves $d(G_t, G_i) = \min_{\boldsymbol{T}} \boldsymbol{T} \odot \boldsymbol{C}$ can be approximated by a differentiable Sinkhorn-Knopp algorithm [337, 338] following [339],

$$\boldsymbol{T} = \text{diag}(\boldsymbol{p}) \exp(-\boldsymbol{C}/\gamma) \text{ diag}(\boldsymbol{q}), \tag{3.18}$$

where $\boldsymbol{p} \in \mathbb{R}_+^{n \times 1}$ and $\boldsymbol{q} \in \mathbb{R}_+^{m \times 1}$. Starting with any positive vector $\boldsymbol{q}^0$ to perform the following iteration:

$$\begin{aligned} &\text{for} \quad i = 0, 1, 2, \ldots \text{ until convergence,} \\ &\quad \boldsymbol{p}^{i+1} = \mathbf{1} \oslash (\boldsymbol{K}\boldsymbol{q}^i), \quad \boldsymbol{q}^{i+1} = \mathbf{1} \oslash (\boldsymbol{K}^\top \boldsymbol{p}^{i+1}), \end{aligned} \tag{3.19}$$

where $\oslash$ denotes element-wise division. $\boldsymbol{K} = \exp(-\boldsymbol{C}/\gamma)$. A computational $\boldsymbol{T}^k$ can be obtained by iterating for a finite number k times,

$$\boldsymbol{T}^k := \text{diag}(\boldsymbol{p}^k)\boldsymbol{K}\text{diag}(\boldsymbol{q}^k). \tag{3.20}$$

3.4.4 Contrastive Learning Objective

We optimize the cosine similarity between image i and positive description t^+ to be close to 1, while negative descriptions t^- to be close to 0,

$$L_1 = \sum\nolimits_{\langle t,i \rangle} D_{KL}(s(t,i) \;||\; \mathbb{1}_{t \in T^+}), \tag{3.21}$$

where $D_{KL}(\cdot||\cdot)$ is the Kullback-Leibler divergence, and $\mathbb{1}_{t \in T^+}$ is the indicator function showing whether the description is a positive description. It enables our model to handle any number of positive and negative descriptions. Also, we include the descriptions of other images in the same batch as negative descriptions.

We also minimize the distance between two event graphs,

$$L_2 = \sum\nolimits_{\langle t,i \rangle} d(G_t, G_i). \tag{3.22}$$

The contrastive learning of event and argument description and the alignment of event graphs are jointly optimized:

$$L = \lambda_1 L_1 + \lambda_2 L_2. \tag{3.23}$$

We set λ_1 and λ_2 as 1 in this paper.

3.5 EXPERIMENTS

3.5.1 Supervised Multimodal Event Extraction

We conduct evaluation on text-only, image-only, and multimodal event mentions in M^2E^2 dataset in Table 3.3. We adopt the traditional event extraction measures, i.e., *Precision*, *Recall* and F_1. For text-only event mentions, we follow [147, 159]: a textual event mention is correct if its event type and trigger offsets match a reference trigger; and a textual event argument is correct if its event type, offsets, and role label match a reference argument. We make a similar definition for image-only event mentions: a visual event mention is correct if its event type and image match a reference visual event mention; and a visual event argument is correct if its event type, localization, and role label match a reference argument. A visual argument is correctly localized if the Intersection over Union (IoU) of the predicted bounding box with the ground truth box of the same role is over 0.5. Finally, we define a multimodal event mention to be correct if its event type and trigger offsets (or the image) match the reference trigger (or the reference image). The arguments of multimodal events are either text or visual arguments, and are evaluated accordingly. To generate bounding boxes for the attention-based model, we threshold the heatmap using the adaptive value of $0.75 * p$, where p is the peak value of the heatmap. Then we compute the tightest bounding box that encloses all of the thresholded region.

Baselines The baselines include:

(1) **Text-only** models: We use the state-of-the-art model JMEE [156] and GAIL [155] for comparison. We also evaluate the effectiveness of cross media joint training by including a version of our model trained only on ACE, denoted as $\text{WASE}^{\mathbb{T}}$.

(2) **Image-only** models: Since we are the first to extract newsworthy events, and the most similar work *situation recognition* can not localize arguments in images, we use our model trained only on image corpus as baselines. Our visual branch has two versions, object-based and attention-based, denoted as $\text{WASE}^{\mathbb{I}}{}_{\text{obj}}$ and $\text{WASE}^{\mathbb{I}}{}_{\text{att}}$.

(3) **Multimodal** models: To show the effectiveness of structured embedding, we include a baseline by removing the text and image GCNs from our model, which is denoted as Flat. The Flat baseline ignores edges and treats images and sentences as sets of vectors. We also compare to the state-of-the-art cross-media common representation model, Contrastive Visual Semantic Embedding VSE-C [340], by training it the same way as WASE.

Training	Model	Text-Only Evaluation					
		Event Mention			Argument Role		
		P	R	F_1	P	R	F_1
Text	JMEE	42.5	58.2	48.7	22.9	28.3	25.3
	GAIL	43.4	53.5	47.9	23.6	29.2	26.1
	WASE$^{\mathbb{T}}$	42.3	58.4	48.2	21.4	30.1	24.9
Multimodal	VSE-C	33.5	47.8	39.4	16.6	24.7	19.8
	Flat$_{\text{att}}$	34.2	63.2	44.4	20.1	27.1	23.1
	Flat$_{\text{obj}}$	38.3	57.9	46.1	21.8	26.6	24.0
	WASE$_{\text{att}}$	37.6	66.8	48.1	27.5	33.2	**30.1**
	WASE$_{\text{obj}}$	42.8	61.9	**50.6**	23.5	30.3	26.4

Table 3.7: Event and argument extraction results (%) on text-only evaluation.

Training	Model	Image-Only Evaluation					
		Event Mention			Argument Role		
		P	R	F_1	P	R	F_1
Image	WASE$^{\mathbb{I}}{}_{\text{att}}$	29.7	61.9	40.1	9.1	10.2	9.6
	WASE$^{\mathbb{I}}{}_{\text{obj}}$	28.6	59.2	38.7	13.3	9.8	11.2
Multimodal	VSE-C	30.3	48.9	26.4	5.6	6.1	5.7
	Flat$_{\text{att}}$	27.1	57.3	36.7	4.3	8.9	5.8
	Flat$_{\text{obj}}$	26.4	55.8	35.8	9.1	6.5	7.6
	WASE$_{\text{att}}$	32.3	63.4	42.8	9.7	11.1	10.3
	WASE$_{\text{obj}}$	43.1	59.2	**49.9**	14.5	10.1	**11.9**

Table 3.8: Event and argument extraction results (%) on image-only evaluation.

Training	Model	Multimodal Evaluation					
		Event Mention			Argument Role		
		P	R	F_1	P	R	F_1
Text	JMEE	42.1	34.6	38.1	21.1	12.6	15.8
	GAIL	44.0	32.4	37.3	22.7	12.8	16.4
	WASE$^{\mathbb{T}}$	41.2	33.1	36.7	20.1	13.0	15.7
Image	WASE$^{\mathbb{I}}{}_{\text{att}}$	28.3	23.0	25.4	2.9	6.1	3.8
	WASE$^{\mathbb{I}}{}_{\text{obj}}$	26.1	22.4	24.1	4.7	5.0	4.9
Multimodal	VSE-C	33.3	48.2	39.3	11.1	14.9	12.8
	Flat$_{\text{att}}$	33.9	59.8	42.2	12.9	17.6	14.9
	Flat$_{\text{obj}}$	34.1	56.4	42.5	16.3	15.9	16.1
	WASE$_{\text{att}}$	38.2	67.1	49.1	18.6	21.6	**19.9**
	WASE$_{\text{obj}}$	43.0	62.1	**50.8**	19.5	18.9	19.2

Table 3.9: Event and argument extraction results (%) on multimodal evaluation.

Quantitative Performance As shown in Table 3.7, Table 3.8 and Table 3.9, our complete methods (WASE_{att} and WASE_{obj}) outperform all baselines in the three evaluation settings in terms of F_1. Our model outperforms its text-only and image-only variants on multimodal events, showing the inadequacy of single-modal information for complex news understanding. Furthermore, our model achieves better performance on text-only and image-only events, which demonstrates the effectiveness of multimodal training framework in knowledge transfer between modalities.

WASE_{obj} and WASE_{att}, are both superior to the state of the art and each has its own advantages. WASE_{obj} predicts more accurate bounding boxes since it is based on a Faster R-CNN pretrained on bounding box annotations, resulting in a higher argument precision. While WASE_{att} achieves a higher argument recall as it is not limited by the predefined object classes of the Faster R-CNN.

Model	P (%)	R (%)	F_1 (%)
rule_based	10.1	100	18.2
VSE	31.2	74.5	44.0
Flat_{att}	33.1	73.5	45.6
Flat_{obj}	34.3	76.4	47.3
WASE_{att}	39.5	73.5	51.5
WASE_{obj}	40.1	75.4	52.4

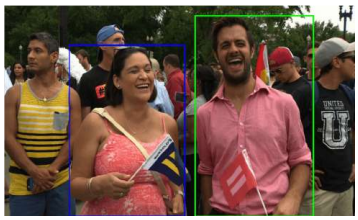
Table 3.10: Cross-media event coreference performance.

Furthermore, to evaluate the cross-media event coreference performance, we pair textual and visual event mentions in the same document, and calculate *Precision*, *Recall* and F_1 to compare with ground truth event mention pairs[8]. As shown in Table 3.10, WASE_{obj} outperforms all multimodal embedding models, as well as the rule-based baseline using event type matching. This demonstrates the effectiveness of our cross-media soft alignment.

Qualitative Analysis Our cross-media joint training approach successfully boosts both event extraction and argument role labeling performance. For example, in Figure 3.10(a) (a), the text-only model can not extract Justice.Arrest event, but the joint model can use the image as background to detect the event type. In Figure 3.10(a) (b), the image-only model detects the image as Conflict.Demonstration, but the sentences in the same document help our model not to label it as Conflict.Demonstration. Compared with multimodal flat embedding in Figure 3.10(b), WASE can learn structures such as Artifact

[8]We do not use coreference clustering metrics because we only focus on mention-level cross-media event coreference instead of the full coreference in all documents.

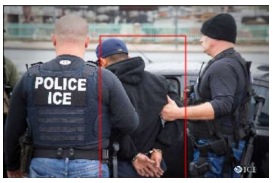
Iraqi security forces *search* ***[Justice.Arrest]*** a civilian in the city of Mosul.



People celebrate Supreme Court ruling on Same Sex Marriage in front of the Supreme Court in Washington.

(a) Image helps textual event extraction, and surrounding sentence helps visual event extraction.

Flat	Event	Movement.Transport
	Role	Artifact = none
Ours	Event	Movement.Transport
	Role	Artifact = man



Flat	Event	Justice:ArrestJail
	Role	Agent = man
Ours	Event	Conflict.Attack
	Role	Entity = man

(b) Comparison with multimodal flat embedding.

Figure 3.10: Architecture for evaluation tasks.

is on top of VEHICLE, and the person in the middle of JUSTICE.ARREST is ENTITY instead of AGENT.

3.5.2 Zero-Shot Multimodal Event Extraction

Setting	Model	Multimedia Event Extraction					
		Event			Argument		
		P	R	F_1	P	R	F_1
Zero-shot	CLIP	29.5	65.7	40.7	9.2	12.7	10.7
	CLIP pretrained on news	31.7	64.7	42.6	9.7	13.1	11.1
	CLIP-Event	36.4	70.8	**48.1**	13.9	16.0	**14.8**
	w/o OptimalTransport	35.0	59.3	44.1	11.0	12.6	11.9
	Single Template	32.3	71.4	44.4	11.9	15.6	13.2
	Composed Template	33.9	72.8	46.3	12.7	15.3	13.9
	Continuous Prompt	33.6	75.7	46.5	11.1	16.7	13.3
	Caption Editing	30.9	71.4	43.2	11.6	13.8	12.6
	GPT-3 Prompt	34.2	76.5	47.3	12.1	16.8	14.1
Supervised	State-of-the-Art [4]	43.1	59.2	49.9	14.5	10.1	11.9
	CLIP finetuned on SWiG	38.1	71.6	49.8	20.9	12.8	15.9
	CLIP-Event$^{+\text{SWiG}}$	41.3	72.8	**52.7**	21.1	13.1	**17.1**
	w/o OptimalTransport	40.3	71.3	51.5	20.8	13.0	16.0

Table 3.11: Evaluation results and ablation studies on Multimedia Event Extraction (M^2E^2).

Baselines We include the following baselines:

Setting	Model	Grounded Situation Recognition (SWiG)				
		Event	Argument			
		verb	value	value-all	ground	ground-all
Zero-shot	CLIP	28.3	13.3	7.6	11.2	3.8
	CLIP pretrained on news	29.9	14.0	8.2	12.0	4.3
	CLIP-Event	**31.4**	**14.9**	**9.2**	**12.8**	**5.2**
	w/o OptimalTransport	30.2	14.2	8.4	12.3	4.4
	Single Template	30.4	14.4	8.6	12.4	4.7
	Composed Template	30.9	14.5	8.8	12.4	4.8
	Continuous Prompt	30.4	14.0	8.3	12.1	4.3
	Caption Editing	30.1	13.9	8.2	12.3	4.4
	GPT-3 Prompt	31.1	14.9	9.1	12.7	5.2
Supervised	State-of-the-Art [187]	39.9	31.4	18.9	24.9	9.7
	CLIP finetuned on SWiG	42.6	32.6	19.2	25.2	10.2
	CLIP-Event$^{+\text{SWiG}}$	**45.6**	**33.1**	**20.1**	**26.1**	**10.6**
	w/o OptimalTransport	44.7	32.9	19.4	24.4	10.1

Table 3.12: Evaluation results and ablation studies on Grounded Situation Recognition (SWiG).

Model	Flickr30k		MSCOCO		VOANews	
CLIP	62.2	81.9	30.3	50.3	21.2	23.4
CLIP pretrained on news	64.3	81.2	32.2	50.8	23.5	25.1
CLIP-Event	**67.0**	**82.6**	**34.0**	**51.3**	**27.5**	**28.7**
w/o OptimalTransport	65.6	80.5	32.5	51.0	25.5	26.9

Table 3.13: R@1(%) on text-to-image (left) and image-to-text (right) retrieval.

Model	VCR		VisualCOMET
	Answer F_1	Rationale F_1	Accuracy@50
Perplexity in [341]	-	-	18.2
CLIP	51.1	46.8	20.1
CLIP pretrained on news	51.8	47.2	20.9
CLIP-Event	**52.4**	**49.2**	**22.4**
w/o OptimalTransport	52.0	48.6	21.1

Table 3.14: Results (%) on zero-shot VCR and VisualCOMET.

(1) **State-of-the-art Multimodal Pretraining Models.** We compare with CLIP [30] by running the public release of "ViT-B/32" and report the scores in the following experiments for a fair comparison. We further pretrain CLIP using the image-captions in the same

dataset in Table 3.5 for a fair comparison in terms of data resources.

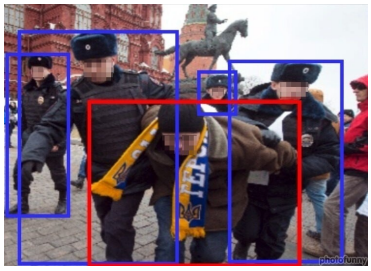
(2) **State-of-the-art Event Extraction Models.** The state-of-the-art event extraction models, such as WASE [4] for Multimodal Event Extraction task, JSL [187] for Grounded Situation Recognition task.

(3) **Ablation Study: CLIP-Event w/o Optimal Transport** is included as a variant of our model in which we remove the alignment between event graphs. It is trained only on the contrastive loss L_1.

(4) **Ablation Study: Each Prompt Function** is used solely during training, for the purpose of comparing its effectiveness.

Analysis on Event Extraction Tasks Under zero-shot settings, we achieve 5.5% absolute F-score gain on event extraction, and 33.3% relative gain on argument extraction on M^2E^2, as shown in Table 3.11 and Table 3.12. The gains achieved by pretraining on news data are significantly amplified with the help of structural event knowledge. For example, CLIP pretrained on news achieves 1.9% improvement compared to the vanilla CLIP on M^2E^2. Our CLIP-Event significantly boosts the gain to 3.89 times.

Zero-shot CLIP-Event outperforms the state-of-the-art weakly supervised model on argument extraction on M^2E^2dataset, showing that the proposed optimal transport alignment effectively captures the argument structures, which previous vision-language pretraining models fail.



Event Type	Arrest
Agent	person
Detainee	person

Event Type	protesting
Agent	people
Place	outdoors

(a) An example result on M^2E^2. (b) An example result on SWiG.

Figure 3.11: Example results of zero-shot multimodal event extraction.

For argument localization, CLIP-Event achieves a higher gain on M^2E^2 than SWiG, due to the fact that SWiG uses a different argument bounding box grounding strategy. SWiG merges all objects that play the same role into a single large bounding box. As shown in

Figure 3.11(b), our approach detects argument roles for each object first, and then merges those objects of the same role into the a large bounding box. In comparison, M^2E^2allows multiple objects with the same argument role, which is consistent with our approach to use objects aligning with argument roles, as shown in Figure 3.11(a).

Analysis on Downstream Tasks We evaluate the system on the following downstream tasks:

(1) Image Retrieval: (a) VOANews presents a greater challenge due to the various events in the captions and the more difficult sentence structures compared to Flickr30k and MSCOCO, as shown in Figure 3.12. The improvement on VOANews is much higher than the gains on Flickr30k and MSCOCO, proving that our model is capable of handling lengthy sentences, particularly those with many events. (b) Downstream tasks benefit from fine-grained event graph alignments. For example, in Figure 3.12, the strong alignment between objects and *investigators* and *destroyed car* enables the image to be successfully ranked higher.

Figure 3.12: Example results of text-to-image retrieval on VOANews, with the visualizations of the optimal transport plan.

(2) VCR: (a) On VCR, the rationale F_1 improves more than answer F_1. Rationale prediction is more challenging since it refers to the details of the scene to justify the answer prediction, showing that our model can well support understand complicated text information. (b) Event knowledge is especially helping with the reasoning on the downstream tasks.

(3) VisualCOMET: We compare our results to the perplexity of the state-of-the-art model [341], which is also retrieval-based. The baseline is trained using the training set of VisualCOMET, but our model is an unsupervised model, which achieves superior performance, demonstrating that our model is capable of comprehending events in the images.

Ablation Studies We have the following observations:

(1) Effect of Event Graph Alignment via Optimal Transport: (a) Removing optimal transport ("w/o OptimalTransport") generally lowers the performance on all evaluation tasks, since it ignores the event graph structures and their cross-media alignment, but relies solely on the overly simplistic image and sentence features. (b) The performance gain on argument extraction task is the highest, since it requires the fine-grained alignment of text and images. (c) We visualize the transport plan in Figure 3.12 to bring insights into the learned alignment. It is a global decision that takes the argument structures of two event graphs into account. Thus, distinct argument roles tend to be associated with diverse objects with different visual features in order to achieve a low *global* transport cost. For instance, *investigators* match objects dressed in white, but not soldier objects, due to the dissimilar visual features. Additionally, one argument role tends to be aligned with objects that have similar visual features, e.g., two *investigators* are both dressed in white protection suits.

(2) Comparison between prompt functions: As shown in Table 3.11, GPT3 provides the optimal performance among prompt functions. It leverages the knowledge encoded in GPT3, thus generating natural descriptions with precise event information. Other prompt functions also demonstrate their effectiveness in supporting event understanding.

3.6 CONCLUSIONS AND FUTURE WORK

In this chapter, we propose a new task of multimodal event extraction to jointly extract events from text and images, and set up the first benchmark. We first explore the supervised training by developing a novel multimodal structured common space construction method to take advantage of the existing image-caption pairs and single-modal annotated data for weakly supervised training. Then we propose a self-supervised training to tackle zero-shot multimodal event extraction by integrating structural event knowledge into vision-language pretraining. We perform cross-media transfer of event knowledge, by automatically extracting event knowledge from captions and supervising image event structure understanding via contrastive learning. We generate hard negatives by manipulating event structures based on confusion matrices, and design event prompt functions to encode events into natural sentences. To transfer argument structural knowledge, we propose an event graph alignment loss via optimal transport, obtaining a global alignment based on argument structures. It outperforms the state-of-the-art vision-language pretraining models on event extraction and downstream tasks under zero-shot settings. Experiments on our publicly released benchmark demonstrate its effectiveness as a new step towards a semantic understanding of events in

multimodal data.

As we look to the future, our goal is to augment our framework to enable event extraction from videos and other modalties, and to apply the results of this extraction to a variety of multimodal applications, including cross-modal reasoning and inference, as well as temporal dynamics tracking.

The visual knowledge sphere presents a complex set of structures including object existence, properties, and affordances, as well as spatial relations, multi-object interactions, situational understanding, and temporal and causal relationships. One of the major challenges lies in whether the model can learn to parse the visual data into a multi-granular knowledge representation, capable of automatically capturing semantics at any level of granularity.

One unique challenge of vision and speech understanding is the temporal dimension. There can be variable asynchronies in the data from different modalities. Also, the relationship between different modalities might not be linear. For instance, a gesture in a video (like a nod) might correspond to an affirmation that occurs later in the speech. It can be challenging to precisely align events in different modalities, especially when there's ambiguity. For example, a visual event like a person starting to laugh might precede, coincide with, or follow the corresponding sound.

Another area of future research is exploring improved methods to harness the power of the language space to assist the visual world. The language world excels in capturing long-range and complex reasoning, as well as compositional abilities of semantics. On the other hand, the visual world is good at capturing detailed visual details, particularly tracking status changes, which are often overlooked in text due to reporting bias. Therefore, finding effective ways to merge the strengths of these two domains will be a critical focus in our future research efforts.

CHAPTER 4: DYNAMICS: TEMPORAL EVENT SCHEMA INDUCTION

History tends to repeat itself. Event extraction in Chapter 3 enables us to obtain a large number of historical events. These historical events imply global knowledge about *event schemas*, such as the patterns describing event interactions. To discover the common patterns between multiple events over a long period, temporal ordering is an important and critical dimension.

In this chapter, we hypothesize that two events are connected when their entity arguments are co-referential or semantically related, and propose a novel Event Graph Schema [322], where two event types are connected through multiple paths involving entities that play important roles in a coherent story. Further, we integrate the temporal dimension into schema induction [14]. This kind of schemas can guide our understanding and ability to make predictions about future events, along with background knowledge including location-, and participant-specific and temporally ordered event information.

4.1 WHAT IS AN EVENT SCHEMA?

Given a news article, as shown in Figure 4.2, we construct an instance graph for every two event instances from information extraction (IE) results. In this example, instance graph (a) tells the story about Russia deploying troops to attack Ukraine using tanks from Russia; instance graph (b) is about Ukrainian protesters hit police using stones that are being carried to Maidan Square. We learn a path language model to select salient and coherent paths between two event types and merge them into a graph schema.

4.2 A NEW SCHEMA REPRESENTATION: TEMPORAL EVENT GRAPH SCHEMA

We propose Temporal Event Graph Schema, whose complexity comes from the inclusion of multiple atomic events (and their arguments), relations and temporal order. A **complex event schema** can be used to define the typical structure of a particular type of complex event, e.g., *car-bombing*. Figure 4.1 shows an example schema about *car-bombing* with multiple temporal dependencies between events. Namely, the occurrence of one event may depend on multiple events. For example, the ASSEMBLE event happens after buying both the bomb materials and the vehicle. Also, there may be multiple events following an event, such as the multiple consequences of the ATTACK event in Figure 4.1. That is to say, "the future is not one-dimensional". Our automatically induced probabilistic complex event schema can

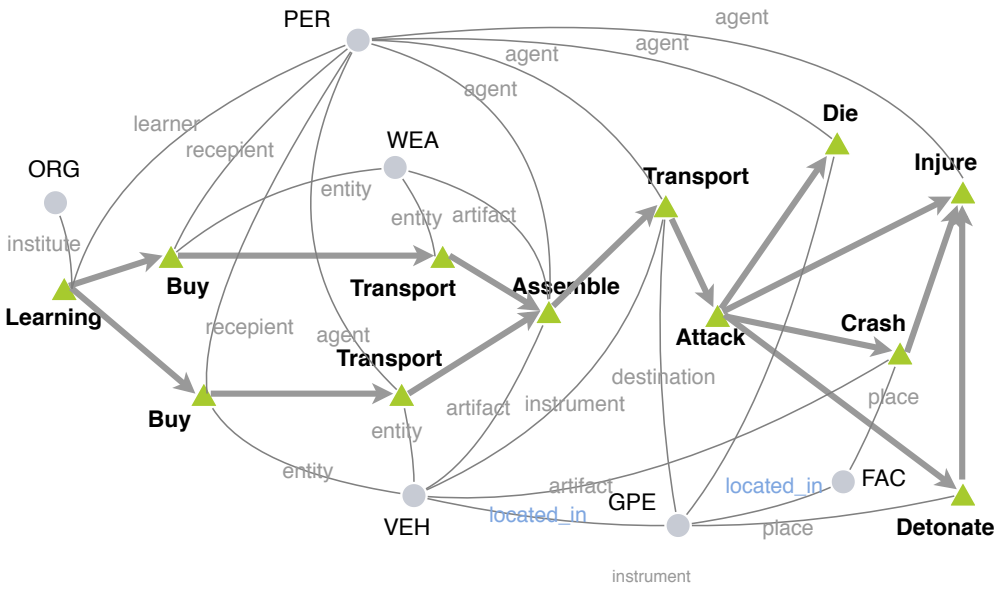


Figure 4.1: The example schema of the complex event type *car-bombing*.

be used to forecast event abstractions into the future and thus provide a comprehensive understanding of evolving situations, events, and trends.

4.3 SCHEMA INDUCTION BETWEEN EVENT PAIRS: PATH LANGUAGE MODELING

4.3.1 Criteria of Event-Event Path Selection

As shown in Table 4.1, a graph schema for two event types consists of **salient** and **coherent** paths between them. A salient path reveals knowledge of recurring event-event connection patterns. For example, the frequent path in Table 4.1 shows that the attacker is a member of the government conducting a deployment, which repeatedly appears in the story about attackers sending weapons and people to attack a target place. However, the attacker is unlikely to be affiliated with a target place, so the infrequent path in Table 4.1 should be excluded from the schema.

In addition, a good path is semantically coherent. For example, the coherent path in Table 4.1 shows that the origin of transportation is a subarea of the attacker's country, which captures the hierarchical part-whole relation between two places. However, in the bad path example, a person is affiliated with both the origin and destination of the transportation, which is a weakly coherent situation.

Furthermore, multiple paths in a good schema should be semantically consistent, namely

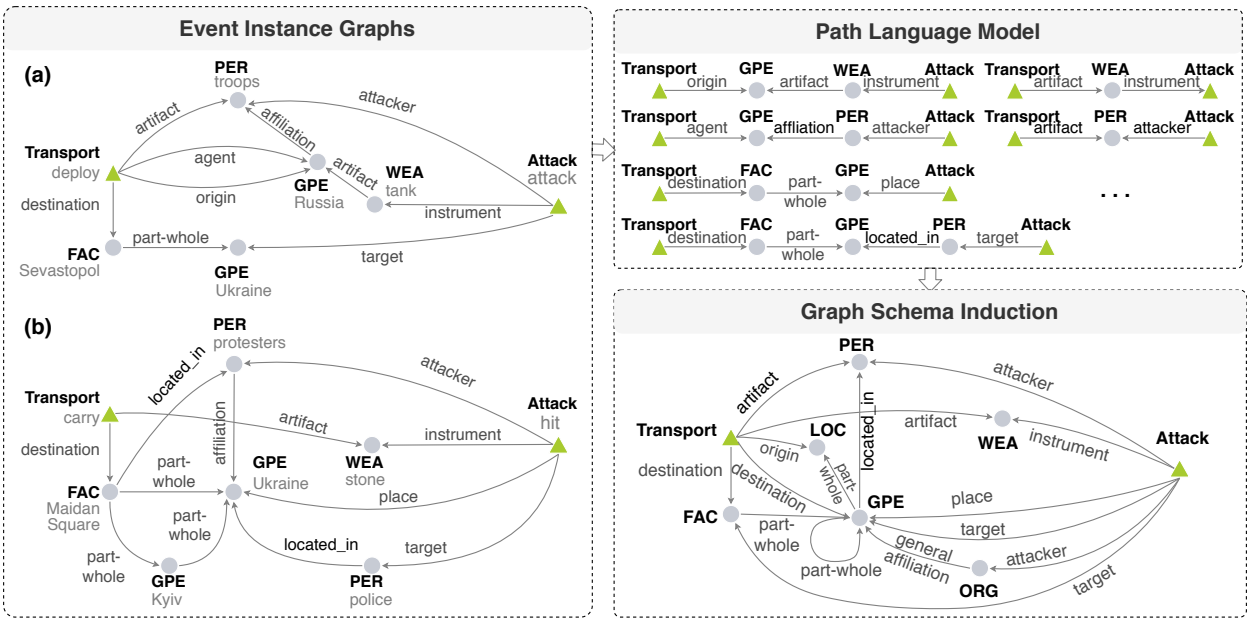


Figure 4.2: The framework of event graph schema induction.

Criteria		Examples	Frequency
Single Path	Salience	**High** TRANSPORT $\xrightarrow{\text{AGENT}}$ GPE $\xrightarrow{\text{AFFILIATION}^{-1}}$ PER $\xrightarrow{\text{ATTACKER}^{-1}}$ ATTACK	31
		Low TRANSPORT $\xrightarrow{\text{DESTINATION}}$ GPE $\xrightarrow{\text{AFFILIATION}^{-1}}$ PER $\xrightarrow{\text{ATTACKER}^{-1}}$ ATTACK	2
	Semantic Coherence	**High** TRANSPORT $\xrightarrow{\text{ORIGIN}}$ FAC $\xrightarrow{\text{PART-WHOLE}}$ LOC $\xrightarrow{\text{PART-WHOLE}}$ GPE $\xrightarrow{\text{AFFILIATION}^{-1}}$ PER $\xrightarrow{\text{ATTACKER}^{-1}}$ ATTACK	9
		Low TRANSPORT $\xrightarrow{\text{AGENT}}$ GPE $\xrightarrow{\text{AFFILIATION}^{-1}}$ PER $\xrightarrow{\text{AFFILIATION}}$ GPE $\xrightarrow{\text{RESIDENT}^{-1}}$ PER $\xrightarrow{\text{TARGET}^{-1}}$ ATTACK	24
Multiple Paths	Semantic Consistency	**High** TRANSPORT $\xrightarrow{\text{DESTINATION}}$ GPE $\xrightarrow{\text{PLACE}^{-1}}$ ATTACK TRANSPORT $\xrightarrow{\text{ARTIFACT}}$ PER $\xrightarrow{\text{LOCATED\_IN}}$ GPE $\xrightarrow{\text{PLACE}^{-1}}$ ATTACK	20
		Low TRANSPORT $\xrightarrow{\text{DESTINATION}}$ GPE $\xrightarrow{\text{PLACE}^{-1}}$ ATTACK TRANSPORT $\xrightarrow{\text{ORIGIN}}$ GPE $\xrightarrow{\text{PLACE}^{-1}}$ ATTACK	0

Table 4.1: The criteria of path ranking to construct event schema graph.

they should co-occur frequently in the same scenario. For example, in Table 4.1, the destination of transportation is the attack's target, and meanwhile, is the location of the transported people. The co-occurrence of these two paths represents a repetitive pattern to connect TRANSPORT and ATTACK. However, the incoherent example in Table 4.1 indicates that the

attack place is both the destination and the origin of the transportation, where two paths rarely co-occur.

To induce such salient and coherent graph schemas, we start by applying Information Extraction (IE) to construct instance graphs between event instances in each document. We construct an event instance graph g for two event instances v and v', that includes all *instance paths* between them. Each instance path starting from v and ending at v' ,

$$p^{\mathbb{I}} = [v, e_{0;1}, v_1, \ldots, e_{n-1;n}, v'] \tag{4.1}$$

is a sequence of nodes $v, v_1, ..., v' \in V$ and edges $e_{0;1}, ..., e_{n-1;n} \in E$. An event-event path is a sequence of types of nodes and edges,

$$p = [\varphi(v), \varphi(e_{0;1}), \varphi(v_1), \ldots, \varphi(e_{n-1;n}), \varphi(v')] \,. \tag{4.2}$$

where $\varphi : \{V, E\} \rightarrow \{\Phi, \Psi\}$ is a mapping function to obtain the type of each node or edge. For example, the path abstracted from the instance path above is ATTACK $\xrightarrow{\text{INSTRUMENT}}$ WEA $\xrightarrow{\text{ARTIFACT}}$ GPE $\xrightarrow{\text{AGENT}^{-1}}$ TRANSPORT . We consider paths in both directions.

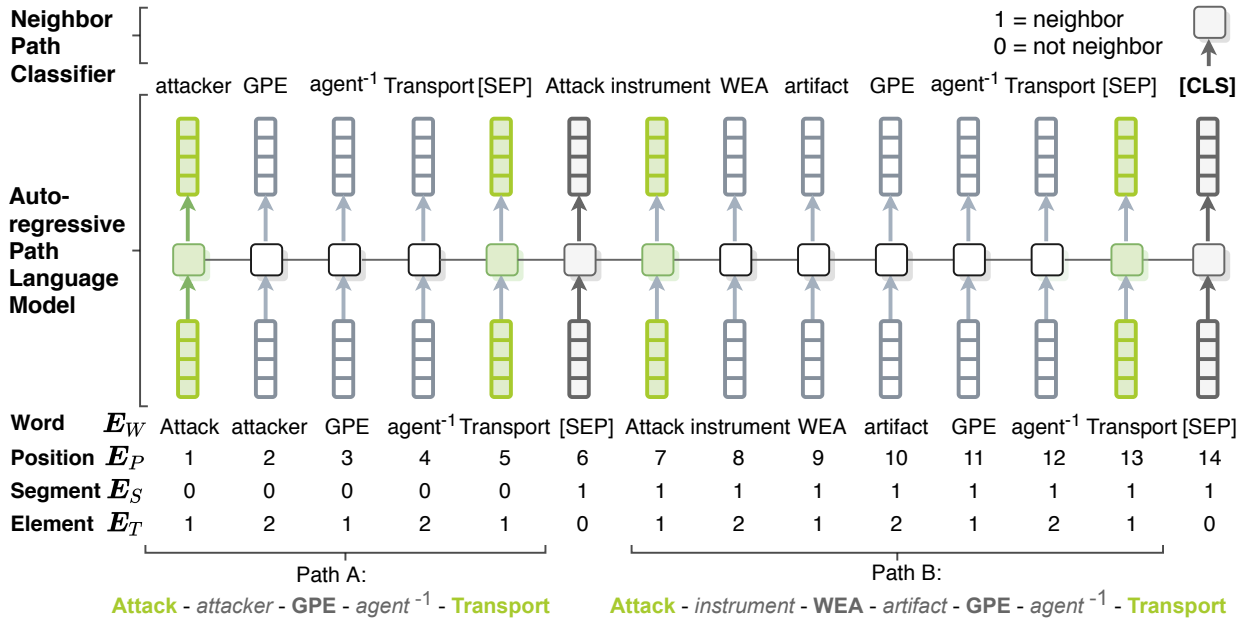


Figure 4.3: Autoregressive path language model with neighbor path classification.

4.3.2 Autoregressive Path Language Model

We consider a path sequence as a text sequence, and learn an auto-regressive path language model to score each path. For a path instance $p^{\mathbb{I}}$, we estimate the probability distribution of a node type $\varphi(v_i)$ (or edge type $\varphi(e_{j;j+1})$), given the sequence of previously observed nodes and edges $[\varphi(v), \varphi(e_{0;1}), \varphi(v_1), \ldots, \varphi(e_{i-1;i})]$, (or $[\varphi(v), \varphi(e_{0;1}), \varphi(v_1), \ldots, \varphi(v_i)]$ if ending with an node), i.e.,

$$\mathcal{L}_{LM} = \sum_{p^{\mathbb{I}}} \Big[\sum_{v_i \in p^{\mathbb{I}}} \log P(\varphi(v_i) | \varphi(v), ..., \varphi(e_{i-1;i})) + \sum_{e_{j;j+1} \in p^{\mathbb{I}}} \log P(\varphi(e_{j;j+1}) | \varphi(v), \varphi(e_{0;1}), ..., \varphi(v_i)) \Big]. \tag{4.3}$$

Following [342], we apply the Transformer [343] to learn the probability distribution, with permutation operation [342] to capture bidirectional contexts. Unlike in text sequences, we have nodes and edges that alternate within path sequences. As shown in Figure 4.3, to distinguish nodes and edges, we add type embedding $\boldsymbol{E}_T = [1, 2, 1, \ldots, 2, 1]$ into the token representation, where 1 stands for nodes, 2 for edges, and 0 for special tokens such as [CLS].

4.3.3 Neighbor Path Classification

To capture the consistency between paths, we train a binary neighbor path classifier to learn the occurrence probability of two paths. For each path $p_i \in \mathcal{P}_{\langle v, v' \rangle}$ between two event instances v and v', we obtain its *neighbor path* set as its co-occuring paths between the same event instances v and v',

$$\mathcal{N}_{p_i} = \{p_j | p_j \in \mathcal{P}_{\langle v, v' \rangle}, v, v' \in V\}. \tag{4.4}$$

We sample negative neighbor paths from paths that appear between the same event types $\varphi(v)$ and $\varphi(v')$, but never occur with p_i in the corpus.

$$\mathcal{N}'_{p_i} = \{p_j | p_j \in \mathcal{P}_{\langle \varphi(v), \varphi(v') \rangle}, p_j \notin \mathcal{N}_{p_i}\}. \tag{4.5}$$

We also swap each path pair to improve the consistency of the neighbor path classification. The neighbor path classifier (top of Figure 4.3) is a linear layer with the classification token $\boldsymbol{x}_{[\text{CLS}]}$ as input,

$$P(p_j \in \mathcal{N}_{p_i}) = \text{sigmoid}\left(\boldsymbol{W}\boldsymbol{x}_{[\text{CLS}]} + \boldsymbol{b}\right). \tag{4.6}$$

We balance the positive and negative path pairs during training, and optimize cross-entropy loss,

$$\mathcal{L}_{NP} = \sum_{p_i} \Big[\sum_{p_j \in \mathcal{N}_{p_i}} \log P(p_j \in \mathcal{N}_{p_i}) + \sum_{p_j \in \mathcal{N}'_{p_i}} \log(1 - P(p_j \in \mathcal{N}_{p_i})) \Big]. \tag{4.7}$$

4.3.4 Joint Training.

We train the path language model on two tasks by jointly optimizing autoregressive language model loss and neighbor path classifier loss,

$$\mathcal{L} = \mathcal{L}_{LM} + \lambda \mathcal{L}_{NP}. \tag{4.8}$$

4.3.5 Graph Schema Construction

. Given two event types ϕ and ϕ', we construct a graph schema s by merging the top k percent ranked paths. Paths in $\mathcal{P}_{\langle \phi, \phi' \rangle}$ are ranked in terms of a score function $f(p)$,

$$f(p_i) = f_{\text{LM}}(p_i) + \alpha f_{\text{NP}}(p_i), \tag{4.9}$$

where $f_{\text{LM}}(p)$ captures salience and coherence of a single path,

$$f_{\text{LM}}(p_i) = \log P([\phi, \psi_{0;1}, \phi_1, \psi_{1;2}, ..., \phi']), \tag{4.10}$$

and where $f_{\text{NP}}(p)$ scores a path p_i by its average probability of co-occuring with other paths $p_j \in \mathcal{P}_{\langle \phi, \phi' \rangle}$ between the given event types ϕ and ϕ',

$$f_{\text{NP}}(p_i) = \frac{1}{|\mathcal{P}_{\langle \phi, \phi' \rangle}|} \sum_{p_j \in \mathcal{P}_{\langle \phi, \phi' \rangle}} \log P(p_j \in \mathcal{N}_{p_i}). \tag{4.11}$$

We merge instance paths into a graph schema s by mapping nodes of the same type into a single node. We allow some self-loops in the graph, such as GPE $\xrightarrow{\text{PART-WHOLE}}$ GPE. Each path in the schema has a probability,

$$P(p_i) = \frac{\exp(f(p_i))}{\sum_{p_j \in s} \exp(f(p_j))}. \tag{4.12}$$

Each edge and node is assigned a salience score by aggregating the scores of paths passing through it,

$$f(\psi_{i;j}) = \sum_{p \in \{p | \psi_{i;j} \in p, p \in s\}} P(p), \quad f(\phi_i) = \sum_{p \in \{p | \phi_i \in p, p \in s\}} P(p). \tag{4.13}$$

4.4 SCHEMA INDUCTION IN EVENT GRAPHS: TEMPORAL EVENT GRAPH MODEL

From a set of documents describing a complex event, we construct an **instance graph** G which contains event nodes E and entity nodes (argument nodes) V. There are three types of edges in this graph:

(1) event-event edges $\langle e_i, e_l \rangle$ connecting events that have direct temporal relations;

(2) event-entity edges $\langle e_i, a, v_j \rangle$ connecting arguments to the event;

(3) entity-entity edges $\langle v_j, r, v_k \rangle$ indicating relations between entities.

We can construct instance graphs by applying Information Extraction (IE) techniques on an input text corpus.

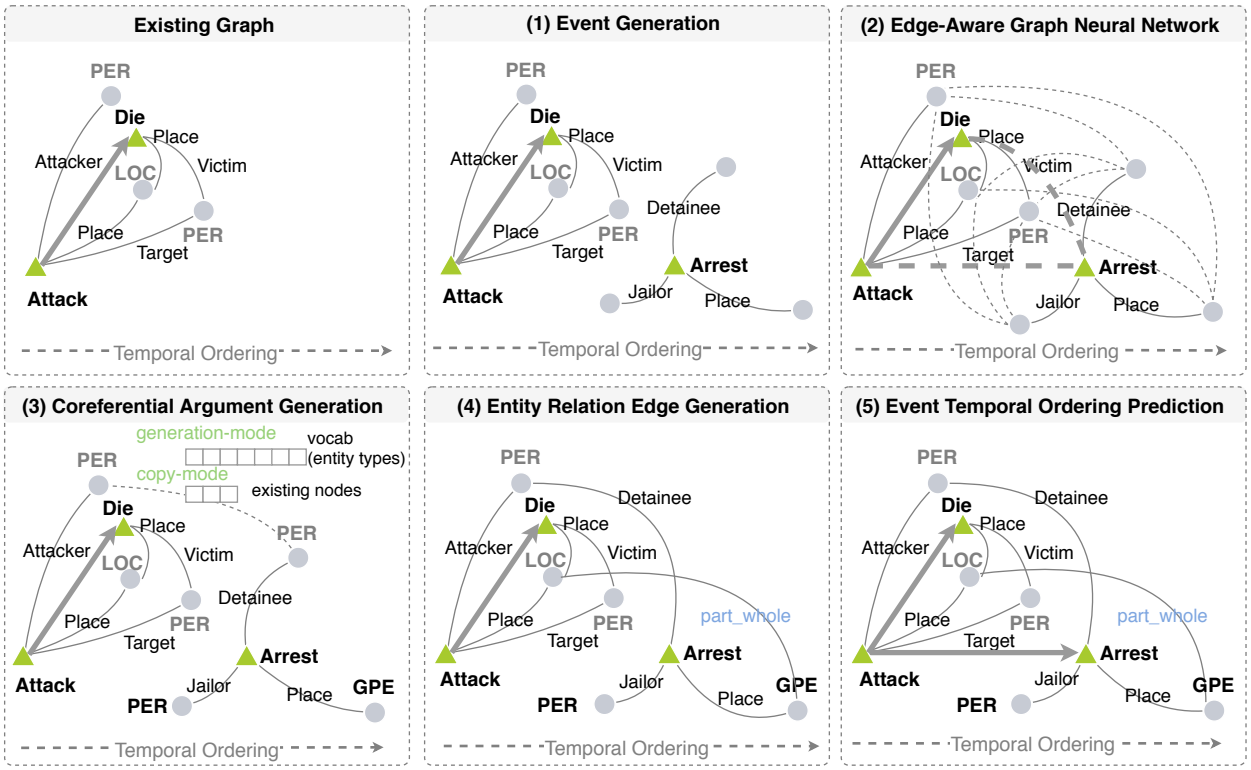


Figure 4.4: The generation process of Temporal Event Graph Model.

4.4.1 Event Graph Generation Overview

Given an instance graph $\mathcal{G}$, we regard the schema as the hidden knowledge to guide the generation of these graphs. To this end, we propose a temporal event graph model that maximizes the probability of each instance graph, parameterized by $\prod_{G \in \mathcal{G}} p(G)$. At each

step, based on the previous graph $G_{<i}$, we predict one event node e_i with its arguments to generate the next graph G_i,

$$p(G) = \prod_{i=0}^{|\mathcal{E}|} p(G_i|G_{<i}). \tag{4.14}$$

We factorize the probability of generating new nodes and edges as:

$$p(G_i|G_{<i}) = p(e_i|G_{<i}) \prod_{a_j \in \mathcal{A}(e_i)} p(\langle e_i, a_j, v_j \rangle | e_i, a_j) \prod_{v_k \in G_{<i}} p(\langle v_j, r, v_k \rangle | v_j, v_k) \prod_{e_l \in G_{<i}} p(\langle e_i, e_l \rangle | e_i, e_l). \tag{4.15}$$

As shown in Figure 4.4, an event node e_i is generated first according to the probability $p(e_i|G_{<i})$. We then add argument nodes based on the IE ontology. We also predict relation $\langle v_j, r, v_k \rangle$ between the newly generated node v_j and the existing nodes $v_k \in G_{<i}$. After knowing the shared and related arguments, we add a final step to predict the temporal relations between the new event e_i and the existing events $e_l \in G_{<i}$.

In the traditional graph generation setting, the order of node generation can be arbitrary. However, in our instance graphs, event nodes are connected through temporal relations. We order events as a directed acyclic graph (DAG). Considering each event may have multiple events both "before" and "after", we obtain the generation order by traversing the graph using Breadth-First Search (BFS).

We also add dummy START/END event nodes to indicate the starting/ending of the graph generation. At the beginning of the generation process, the graph G_0 has a single start event node $e_{[\text{SOG}]}$. We generate $e_{[\text{EOG}]}$ to signal the end of the graph.

4.4.2 The Generation of A Single Event

Event Node Generation To determine the event type of the newly generated event node e_i, we apply a graph pooling over all events to get the current graph representation $\boldsymbol{g}_i$,

$$\boldsymbol{g}_i = \text{Pooling}(\{\boldsymbol{e}_0, \cdots, \boldsymbol{e}_{i-1}\}). \tag{4.16}$$

The event type is predicted through a fully connected layer,

$$p(e_i|G_{<i}) = \frac{\exp(\boldsymbol{W}_{\phi(e_i)} \boldsymbol{g}_i)}{\sum_{\phi' \in \Phi_{\mathcal{E}} \cup [\text{EOG}]} \exp(\boldsymbol{W}_{\phi'} \boldsymbol{g}_i)}. \tag{4.17}$$

Message Passing We use a Graph Neural Network (GNN) [330] to update node embeddings following the graph structure. Before we run the GNN on the graph, we first add *virtual edges* between the newly generated event and all previous events, and between new

entities and previous entities, shown as dashed lines in Figure 4.4. The virtual edges enable the representations of new nodes to aggregate the messages from previous nodes, which has been proven effective in [267].

Coreferential Entity Node Merging After updating the node representations, we detect the entity type of each argument, and also predict whether the argument is coreferential to existing entities. Inspired by copy mechanism [344], we classify each argument node v_j to either a new entity with entity type $\phi(v_j)$, or an existing entity node in the previous graph $G_{<i}$. For example, in Figure 4.4, the DETAINEE should be classified to the existing ATTACKER node, while JAILOR node is classified as PERSON. Namely,

$$p(\langle e_i, a_j, v_j\rangle | e_i, a_j) = \begin{cases} p(\langle e_i, a_j, v_j\rangle, \mathsf{g} | e_i, a_j) & \text{if } v_j \text{ is new,} \\ p(\langle e_i, a_j, v_j\rangle, \mathsf{c} | e_i, a_j) & \text{otherwise,} \end{cases} \tag{4.18}$$

where $p(\langle e_i, a_j, v_j\rangle, \mathsf{g} | e_i, a_j)$ is the generation probability, classifying the new node to its entity type $\phi(v_j)$:

$$p(\langle e_i, a_j, v_j\rangle, \mathsf{g} | e_i, a_j) = \exp(\boldsymbol{W}_{\phi(v_j)} \boldsymbol{v}_j)/Z \tag{4.19}$$

The copy probability $p(\langle e_i, a_j, v_j\rangle, \mathsf{c} | e_i, a_j)$ selects the coreferential entity v from the entities in existing graph, denoted by $V_{<i}$,

$$p(\langle e_i, a_j, v_j\rangle, \mathsf{c} | e_i, a_j) = \exp(\boldsymbol{W}_{v} \boldsymbol{v}_j)/Z. \tag{4.20}$$

Here, Z is the shared normalization term,

$$Z = \sum\nolimits_{\phi' \in \Phi_{\mathcal{V}}} \exp(\boldsymbol{W}_{\phi'} \boldsymbol{v}_j) + \sum\nolimits_{v' \in V_{<i}} \exp(\boldsymbol{W}_{v'} \boldsymbol{v}_j) \tag{4.21}$$

If determined to copy, we merge coreferential entities in the graph.

Entity Relational Edge Generation We model the relation edge generation probability as a categorical distribution over relation types, and add [O] (OTHER) to the typeset $\mathcal{R}$ to represent that there is no relation edge:

$$p(\langle v_j, r, v_k\rangle | v_j, v_k) = \frac{\exp(\text{MLP}_r(\boldsymbol{v}_j - \boldsymbol{v}_k))}{\sum\limits_{r' \in \mathcal{R} \cup [\text{O}]} \exp(\text{MLP}_{r'}(\boldsymbol{v}_j - \boldsymbol{v}_k))} \tag{4.22}$$

We use two hidden layers with ReLU activation functions to implement the MLP.

Event Temporal Ordering Prediction To predict the temporal dependencies between the new events and existing events, we connect them through temporal edges, as shown in Figure 4.4. These edges are critical for message passing in predicting the next event. We build temporal edges in the last phase of generation, since it relies on the shared and related arguments. Considering that temporal edges are interdependent, we model the generation probability as a mixture of Bernoulli distributions following [267]:

$$\begin{aligned} p(\langle e_i, e_l \rangle | e_i, e_l) &= \sum\nolimits_b \gamma_b \; \theta_{b,i,l}, \\ \gamma_b &= \text{Softmax}\left(\sum\nolimits_{i,l} \text{MLP}(\boldsymbol{e}_i - \boldsymbol{e}_l)\right), \\ \theta_{b,i,l} &= \sigma\left(\text{MLP}_\theta(\boldsymbol{e}_i - \boldsymbol{e}_l)\right), \end{aligned} \tag{4.23}$$

where B is the number of mixture components. When $B = 1$, the distribution degenerates to factorized Bernoulli, which assumes the independence of each potential temporal edge conditioned on the existing graph.

4.4.3 Training and Schema Decoding

We optimize the negative log-likelihood loss to learn the generation of each instance event graph G in the historical training graphs $\mathcal{G}_{\text{train}}$,

$$\mathcal{L} = \sum\nolimits_{G \in \mathcal{G}_{\text{train}}} -\log_2 p(G). \tag{4.24}$$

To compose the schema library for each complex event scenario, we construct instance graphs from related documents to learn a graph model, and then obtain the schema using greedy decoding.

4.5 EXPERIMENTS

4.5.1 Evaluation on Path-based Schemas

We use Automatic Content Extraction (ACE) 2005 dataset[9], the widely used dataset with annotated instances of 7 entity types, 6 relation types, 33 event types, and 22 argument roles. We consider the training set as historical data to train the LM, and the test set as our target data to induce schema for target scenarios. Table 4.2 shows the data statistics.

[9] `https://www.ldc.upenn.edu/collaborations/past-projects/ace`

Split	#Documents	#Entities	#Relations	#Events	#Arguments
Historical$_{\text{ann}}$	529	47,525	7,152	4,419	7,888
Historical$_{\text{sys}}$	529	48,664	7,018	4,426	6,614
Validation	40	3,422	728	468	938
Target	30	3,673	802	424	897

Table 4.2: Data statistics.

Instance Coverage We propose Instance Coverage to evaluate whether a graph schema is salient. A salient schema can serve as a skeleton to recover instance graphs. Therefore, we use each graph schema $s \in \mathcal{S}$ to match back to each ground-truth instance graph $g \in \mathcal{G}$ and evaluate their intersection $g \cap s$ in terms of Precision and Recall. Intersection is obtained by searching instance graphs with each graph schema as a query. Since instance graphs can be regarded as *partially* instantiated graph schema, we employ the substructures of the schema graph, i.e., paths of different lengths, as queries. For example, a path of length $l = 3$ is a triple in graph schema $\langle \phi_i, \psi_{ij}, \phi_j \rangle \in s$. We consider an instance triple $\langle v_m, e_{mn}, v_n \rangle \in g$ matched if instance types match, i.e., $\varphi(v_m){=}\phi_i$, $\varphi(e_{mn}){=}\psi_{ij}$, $\varphi(v_n){=}\phi_j$. Let $|\cdot|_{\mathbb{I}}$ denote the number of instance substructures matched, and $|\cdot|_{\mathbb{S}}$ is the number of schema substructures matched, i.e.,

$$\text{Precision} = \frac{\sum_{s\in\mathcal{S}}\sum_{g\in\mathcal{G}}|g\cap s|_{\mathbb{S}}}{\sum_{s\in\mathcal{S}}|s|_{\mathbb{S}}}, \text{Recall} = \frac{\sum_{s\in\mathcal{S}}\sum_{g\in\mathcal{G}}|g\cap s|_{\mathbb{I}}}{\sum_{g\in\mathcal{G}}|g|_{\mathbb{I}}}. \tag{4.25}$$

By extension, each path of length $l{=}5$ in a graph schema $[\phi_i, \psi_{ij}, \phi_j, \psi_{jk}, \phi_k]$ contains two consecutive triples $\langle \phi_i, \psi_{ij}, \phi_j \rangle, \langle \phi_j, \psi_{jk}, \phi_k \rangle \in s$, and a matched instance path contains two consecutive instance triples $\langle v_m, e_{mn}, v_n \rangle, \langle v_n, e_{no}, v_o \rangle \in g$, where $\varphi(v_m){=}\phi_i$, $\varphi(e_{mn}){=}\psi_{ij}$, $\varphi(v_n){=}\phi_j$, $\varphi(e_{no}){=}\psi_{jk}$, $\varphi(v_o){=}\phi_k$. Similarly, a path of length $l{=}7$ contains three consecutive triples.

Instance Coherence For an instance graph between two events v and v', we hypothesize that the graph is coherent if v and v' are from the same discourse (document). We carefully select 24 documents with each document talking about a unique complex event such as *Iraq War* or *North Korea Nuclear Test*. A coherent schema should have the maximal number of matched instance graphs $g \cap s$ from a single document, but the minimal number of matched graphs connecting two event instances from different documents. We define **Instance Coherence** as the proportion of event-event path instances in graphs within one

document.

$$\text{Coherence} = \frac{\sum_{s\in\mathcal{S}}\sum_{g\in\mathcal{G}}\sum_{p\in g\cap s} f(p)\cdot\mathbb{I}_g}{\sum_{s\in\mathcal{S}}\sum_{g\in\mathcal{G}}\sum_{p\in g\cap s} f(p)}, \tag{4.26}$$

where $\mathbb{I}_g$ is an indicator function taking value 1 when g is between event instances from the same document, and value 0 otherwise.

Historical Instance Graphs	**Model**	**Schema@10**								
		$l=7$			$l=5$			$l=3$		
		P	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**
Historical$_{\text{ann}}$	**Frequency**	76.7	9.5	16.9	90.5	48.3	63.0	100	37.5	54.6
	Unigram LM	63.9	7.3	13.1	87.1	35.4	50.3	100	33.7	50.4
	Bigram LM	75.4	8.5	15.3	92.6	36.8	52.6	100	33.4	50.1
	Trigram LM	62.7	8.5	15.0	89.4	41.6	56.7	100	39.9	57.0
	PathLM	54.3	16.6	**25.4**	83.7	63.8	**72.4**	100	41.8	**58.9**
	w/o CLS$_{NP}$	71.2	14.5	24.1	90.3	58.3	70.9	100	39.8	56.9
Historical$_{\text{sys}}$	**Frequency**	68.6	9.8	17.1	87.0	49.4	63.0	100	37.6	54.7
	Unigram LM	54.3	7.5	13.1	83.7	36.2	50.5	100	41.0	58.2
	Bigram LM	61.4	7.9	13.9	88.5	37.7	52.8	100	39.2	56.3
	Trigram LM	65.2	9.8	17.1	89.6	46.8	61.5	100	37.3	54.4
	PathLM	51.8	18.5	**27.3**	83.2	68.0	**74.8**	100	41.7	**58.8**
	w/o CLS$_{NP}$	72.7	14.4	24.1	89.5	55.1	68.2	100	40.1	57.3

Table 4.3: Instance coverage (%) by checking the intersection of schemas@10 and instance graphs.

Results and Analysis We induce 124 and 197 graph schemas for Schema@10 and Schema@20 respectively. Figure 4.2 shows an output graph schema.[10] According to Table 4.3 and Table 4.4, PathLM achieves significant improvement relative to the other methods on instance coverage. Also it outperforms all baselines on instance coherence, as shown in Table 4.5. T-test shows that the gains achieved by PathLM are all statistically significant over baselines (Frequency, UnigramLM, BigramLM, TrigramLM), with a P value less than 0.01.

We make the following observations:

(1) PathLM achieves larger gains compared to baselines on Schema@10 (Table 4.3) than Schema@20 (Table 4.4), demonstrating the effectiveness of our ranking approach, especially on top ranked ones.

(2) The improvement relative to baselines on longer path queries (e.g. $l = 7$) is greater than shorter paths (e.g., $l = 3$) in both Table 4.3 and Table 4.4, showing that our approach

[10]Visualization of schema repository is in `http://blender.cs.illinois.edu/software/pathlm`.

Historical Instance Graphs	Model	Schema@20								
		$l=7$			$l=5$			$l=3$		
		P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
Historical$_{ann}$	**Frequency**	63.6	17.9	28.0	87.6	70.6	78.2	100	42.6	59.7
	Unigram LM	55.4	14.8	23.4	86.0	60.8	71.2	100	43.8	60.9
	Bigram LM	62.6	16.4	26.0	88.1	63.2	73.6	100	43.2	60.3
	Trigram LM	53.4	17.8	26.7	85.6	68.2	75.9	100	44.6	61.6
	PathLM	53.8	27.2	**36.1**	83.0	80.0	81.5	100	44.7	**61.8**
	w/o CLS_{NP}	57.8	25.8	35.6	85.7	80.1	**82.8**	100	42.9	60.1
Historical$_{sys}$	**Frequency**	67.8	19.3	29.9	88.5	70.1	78.2	100	41.6	58.8
	Unigram LM	52.4	17.9	26.7	83.0	66.4	73.8	100	44.6	61.7
	Bigram LM	58.3	15.3	24.2	86.5	63.8	73.4	100	43.5	60.6
	Trigram LM	54.5	17.6	26.6	86.2	68.7	76.5	100	44.1	61.2
	PathLM	49.6	29.3	**36.9**	81.7	85.4	**83.5**	100	44.8	**61.9**
	w/o CLS_{NP}	54.8	24.7	34.0	83.8	75.9	80.0	100	44.7	61.7

Table 4.4: Instance coverage (%) by checking the intersection of schemas@20 and instance graphs.

Historical	Model	Schema@10	Schema@20
Historical$_{ann}$	**Frequency**	67.8	65.6
	Unigram LM	62.4	69.9
	Bigram LM	59.0	67.5
	Trigram LM	56.6	64.9
	PathLM	**76.0**	**79.9**
	w/o CLS_{NP}	75.3	79.2
Historical$_{sys}$	**Frequency**	60.1	65.6
	Unigram LM	61.8	70.0
	Bigram LM	59.7	69.6
	Trigram LM	55.8	65.8
	PathLM	**76.4**	**78.5**
	w/o CLS_{NP}	73.9	77.1

Table 4.5: Instance coherence (%) of schema graphs covering top k percent paths, $k = 10, 20$.

is able to capture complex graph structures involving long distance between related events. In the l=3 setting, the performance of PathLM is close to baselines. The reason is that l=3 setting evaluates a single overlapped triple, which is exactly the objective of TrigramLM. We conduct t-test, and the gain is statistically significant (P value less than 0.01).

(3) The neighbor path classification proves to be effective in enhancing the salience (see 'w/o CLS_{NP}' in Table 4.3 and Table 4.4) and coherence (see 'w/o CLS_{NP}' in Table 4.5) of the induced schemas, showing that salient substructures can be better captured by frequently co-occurring paths. The model outputs consistent neighbor path classification results for the swapped path pairs. 96.17% swapped path pairs yield the same results as original pairs.

(4) The schemas induced from $\text{Historical}_{\text{sys}}$ and $\text{Historical}_{\text{ann}}$ have comparable performance. This proves our approach is robust to extraction noise and effective even with lower quality input.

4.5.2 Evaluation on Event Graph Schemas

We conduct experiments on two datasets for both the general scenario and a more specific scenario. We adopt the DARPA KAIROS[11] ontology, a newly defined fine-grained ontology for Schema Learning, with 24 entity types, 46 relation types, 67 event types, and 85 argument roles. We perform evaluations on two datasets: (1) The Schema Learning Corpus, released by LDC (LDC2020E25), includes 82 types of complex events, such as *Disease Outbreak*, *Presentations* and *Shop Online*. (2) We chose the *improvised explosive device (IED)* as our case study, and collect a dataset from Wikipedia describing 4 types of complex events, i.e., *Car-bombing IED*, *Drone Strikes IED*, *Suicide IED* and *General IED*. Statistics are shown in Table 4.6.

Dataset	Split	#doc	#graph	#event	#arg	#rel
General	**Train**	451	451	6,040	10,720	6,858
	Dev	83	83	1,044	1,762	1,112
	Test	83	83	1,211	2,112	1,363
IED	**Train**	5,247	343	41,672	136,894	122,846
	Dev	575	42	4,661	15,404	13,320
	Test	577	45	5,089	16,721	14,054

Table 4.6: Data statistics. Each instance graph is about one complex event.

Schema Quality Evaluation We compare the generated schemas with the ground truth schemas based on the overlap between them. The following evaluation metrics were employed:[12]

[11] `https://github.com/NextCenturyCorporation/kairos-pub/tree/master/data-format/ontology`

[12] We cannot use graph matching to compare between baselines and our approach due to the difference in the graph structures being modeled.

(1) **Event Match:** A good schema must contain the events crucial to the complex event scenario. *F-score* is used to compute the overlap of event nodes.

(2) **Event Sequence Match:** A good schema is able to track events through a timeline. So we obtain event sequences following temporal order, and evaluate *F-score* on the overlapping sequences of lengths $l = 2$ and $l = 3$.

(3) **Event Argument Connection Match:** Our complex event graph schema includes entities and their relations and captures how events are connected through arguments, in addition to their temporal order. We categorize these connections into three categories: (1) two events are connected by shared arguments; (2) two events have related arguments, i.e., their arguments are connected through entity relations; (3) there are no direct connections between two events. For every pair of overlapped events, we calculate *F-score* based on whether these connections are predicted correctly.

(4) **Instance Graph Perplexity Evaluation** To evaluate our temporal event graph model, we compute the *instance graph perplexity* by predicting the instance graphs in the test set,

$$\text{PP} = 2^{-\frac{1}{|\mathcal{G}_{\text{test}}|}\sum_{G\in\mathcal{G}_{\text{test}}}\log_2 p(G)}. \tag{4.27}$$

We calculate the *full perplexity* for the entire graph using Equation 4.15, and *event perplexity* using only event nodes, emphasizing the importance of correctly predicting events.

Overview of Results In Table 4.7, the significant gain on *event match* demonstrates the ability of our graph model to keep salient events. On *sequence match*, our approach achieves larger performance gain compared to baselines when the path length l is longer. It implies that the proposed model is capable of capturing longer and wider temporal dependencies. In the case of *connection match*, only sequential pattern mining in the baselines can predict connections between events. When compared against sequential pattern mining, our generation model significantly performs better since it considers the inter-dependency of arguments and encodes them with graph structures.

Ablation Study Removing argument generation ("w/o ArgumentGeneration") generally lowers the performance on all evaluation tasks, since it ignores the coreferential arguments and their relations, but relies solely on the overly simplistic temporal order to connect events. This is especially apparent from the instance graph perplexity in Table 4.7.

Learning Corpus Size An average of 113 instance graphs is used for each complex event type in the IED scenario, and 383 instance graphs to learn the schema model in the General

Dataset	Model	Event Match	Sequence Match $l=2$	Sequence Match $l=3$	Connection Match	Event Perplexity	Full Perplexity
General	Event Language Model	54.76	22.87	8.61	-	-	-
	Sequential Pattern Mining	49.18	20.31	7.37	-	-	-
	Event Graph Model	**58.15**	**24.79**	**9.18**	-	**24.25**	137.18
	w/o ArgumentGeneration	56.96	22.47	8.21	-	68.59	-
IED	Event Language Model	49.15	17.77	5.32	-	-	-
	Sequential Pattern Mining	47.91	18.39	4.79	5.41	-	-
	Event Graph Model	**59.73**	**21.51**	**7.81**	**10.67**	**39.39**	168.89
	w/o ArgumentGeneration	55.01	18.24	6.67	-	51.98	-

Table 4.7: Intrinsic evaluation results, including schema matching F1 score (%) and instance graph perplexity.

scenario. The better performance on the IED dataset in Table 4.7 shows that the number of instance graphs increases the schema induction performance.

Effect of Information Extraction Errors Based on the error analysis for schemas induced in Table 1, the effect of extraction errors can be categorized into: (1) temporal ordering errors: 43.3%; (2) missing events: 34.4%; (3) missing coreferential events: 8.8%; (4) incorrect event type: 7.7%; (5) missing coreferential arguments: 5.5%. However, even on automatically extracted event graphs with extraction errors, our model significantly performs better on event prediction compared to human-constructed schemas, as shown in Table 4. It demonstrates that our schema induction method is robust and effective to support downstream tasks, even when only provided with noisy data with extraction errors.

4.6 APPLICATION: SCHEMA-GUIDED INFORMATION EXTRACTION

4.6.1 Settings

As a case study for extrinsic evaluation, we evaluate the impact of our induced schema[13] on end-to-end Information Extraction (IE). We choose the IE system OneIE [345] [14] as our baseline for two reasons: (1) it achieves state-of-the-art performance on all IE components; (2) it can easily incorporate global features during decoding converting each input sentence

[13]The schema is induced from annotated instance graphs of historical data, which is the training data of IE system.

[14]Code is public available at `http://blender.cs.illinois.edu/software/oneie/`

into an instance graph.

Given an input sentence, OneIE generates a set of candidate IE graphs at each decoding step, as shown in Figure 4.5. The candidate IE graphs are ranked by type prediction scores $s'(G)$ of each entity, relation and event in each graph G. We consider schemas as global features and use them as an additional scoring mechanism for OneIE [15]. The schemas are induced from the training data of our IE system. If a path p_i in the schema appears n_i times in a candidate graph, we add $n_i * w_i$ to obtain the global score of this graph,

$$s(G) = s'(G) + \sum_{p_i \in s, s \in \mathcal{S}} n_i * w_i, \tag{4.28}$$

where w_i is a learnable weight. The candidate graphs are then ranked in terms of their global scores. In this way, the model can promote candidate graphs containing positive global features, even if the graphs may have lower local type prediction scores.

4.6.2 Results and Analysis

Model	**Entity**	**Relation**	**Event**			
			Trigger-I	**Trigger-C**	**Argument-I**	**Argument-C**
OneIE Baseline	**90.3**	44.7	75.8	72.7	57.8	55.5
+PathLM	90.2	**60.9**	76.0	**73.4**	59.0	**56.6**
w/o CLS_{NP}	90.1	60.3	75.7	72.8	58.3	55.8

Table 4.8: F_1 score (%) of schema-guided information extraction.

As shown in Table 4.8, our event graph schemas have provided significant improvement on relation extraction and event extraction which require knowledge of complex connections among events and entities. Our approach achieves dramatic improvement on relation extraction, because existing methods mainly rely on local contexts between two entities, which are typically short and ambiguous. In contrast, the paths in our graph schemas can capture the global context between two events, and thus event-related information captures deeper contextual features, yielding a big boost in performance. For example, when decoding candidate IE graph in Figure 4.5, the LOCATED_IN relation is extracted by promoting the structures matching paths in the graph schema.

[15] To show the effectiveness of schema, we remove the original human-designed global features in OneIE.

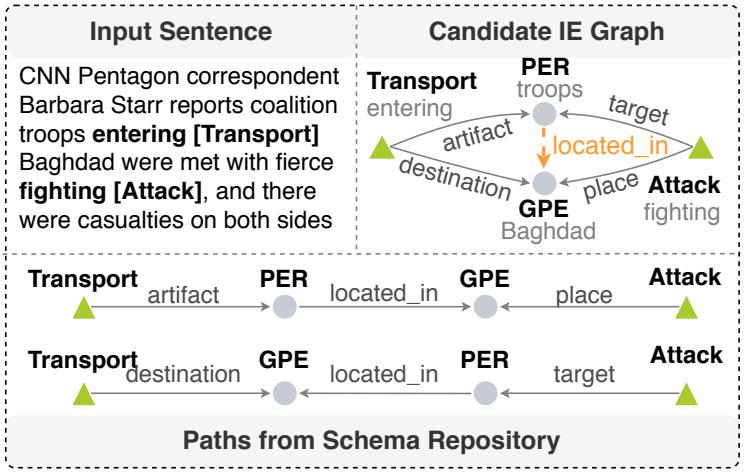


Figure 4.5: An example showing how schema improves the quality of IE by promoting the candidate IE graph matching paths from schema.

4.7 APPLICATION: SCHEMA-GUIDED EVENT PREDICTION

To explore schema-guided probabilistic reasoning and prediction, we perform an extrinsic evaluation of event prediction. Different from traditional event prediction tasks, the temporal event graphs contain arguments with relations, and there are type labels assigned to nodes and edges. We create a graph-based event prediction dataset using our testing graphs. The task aims to predict ending events of each graph, i.e., events that have no future events after it. An event is predicted correctly if its event type matches one of the ending events in the graph. Considering that there can be multiple ending events in one instance graph, we rank event type prediction scores and adopt *MRR (Mean Reciprocal Rank)* and *HITS@1* as evaluation metrics.

Our graph model obtains significant improvement (see Table 4.9.) The low performance of human schema demonstrates the importance of probabilistically modeling schemas to support downstream tasks. Take Figure 4.6 as an example. Human schemas produce incorrect event types such as TrailHearing, since it matches the sequence Attack → Die → Trail-Hearing, incapable of capturing the inter-dependencies between sequences. However, our model is able to customize the prediction to the global context of the input graph, and take into account that there is no Arrest event or justice-related events in the input graph. Also, the human schema fails to predict Injure and Attack, because it relies on the exact match of event sequences of lengths $l \geq 2$, and cannot handle the variants of sequences. This problem can be solved by our probabilistic schema, via modeling the prediction probability conditioned on the existing graph. For example, even though Attack mostly happens before Die, we learn that Attack might repeat after Die event if there are multiple Attack

Dataset	Model	MRR	HITS@1
General	Event Language Model	0.367	0.497
	Sequential Pattern Mining	0.330	0.478
	Human Schema	0.173	0.205
	Event Graph Model	**0.401**	**0.520**
	w/o ArgumentGeneraion	0.392	0.509
IED	Event Language Model	0.169	0.513
	Sequential Pattern Mining	0.138	0.378
	Human Schema	0.072	0.222
	Event Graph Model	**0.224**	**0.741**
	w/o ArgumentGeneraion	0.210	0.734

Table 4.9: Schema-guided event prediction performance.

and DETONATE in the existing graph, which means the complex event is about a series of conflict events.

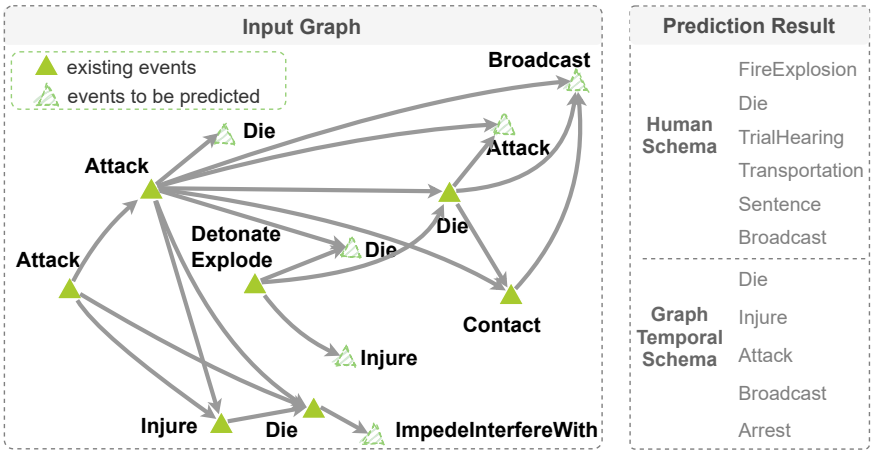


Figure 4.6: An event prediction example (IED scenario).

4.8 CONCLUSIONS AND FUTURE WORK

In this chapter, we propose Event Graph Schema induction as a new step towards the semantic understanding of inter-event connections. We learn knowledge of recurring event interaction patterns by proposing a path language model based method, which is able to construct probabilistic graph schemas containing salient and semantically coherent event-event paths, which also effectively enhances end-to-end Information Extraction. Further, we

extends the path-based induction to graph-based induction, and propose a graph generative model to induce *Temporal Complex Event Schemas*, which are capable of representing multiple temporal dependencies between events and their connected arguments. We induce such schemas by learning an *event graph model*, a deep auto-regressive model, from the automatically extracted instance graphs. Event prediction proves its ability to make predictions with respect to what might happen next, along with background knowledge including location-, and participant-specific and temporally ordered event information.

In the future, we aspire to incorporate logic into our schema induction process. This includes handling interchangeable relationships, optional events, and more complex logic elements such as conditions based on 'and/or' structures. This enhancement will make our models more robust and versatile, able to accurately capture and represent intricate event relationships.

In addition, we aim to ensure that our predictions are accompanied by supporting evidence, thereby enhancing the credibility of the system. This is crucial as it not only increases the trustworthiness of the predictions made but also offers insights into the underlying reasoning, leading to a more reliable and transparent system. Also, we plan to incorporate the instance-level information during schema induction, and further support event prediction to predict instances. One potential way is to explore the graph alignment between schema graphs and instance graphs, and use schema knowledge to improve information extraction.

Alongside these developments, we plan to extend our graph schemas to support rich ontologies in the open domain with hierarchical structures. By assembling our graph schemas to represent more complex scenarios involving multiple events, we seek to apply them to a broader spectrum of downstream applications, including forward/backward event graph completion and event prediction. This comprehensive approach aims to foster a more holistic understanding of events and their intricate relationships.

CHAPTER 5: FACTUALITY: FACT-BASED GENERATION

A key bottleneck in analyzing large corpora is the ability to encode factual knowledge and control the generation process so that the output is factually consistent, with the capability to trace back to the original factual information. Whereas existing studies have built text graphs by augmenting text sequences with different hidden structural information, they are typically entity-centric and overlook the events' intra-structures (arguments) and inter-structures (event-event connections).

In this chapter, we use event graphs to provide a new comprehensive representation and necessary inductive bias. Our goal is to define the multi-document joint representation as the contextualized embeddings of the nodes on the event graph and collectively model events and arguments. These event graphs can then be used to address the massive unstructured data challenge in real-world applications: (1) **Timeline Summarization** [20, 21] is formulated as an event graph compression problem and then I design time-aware optimal transport to obtain the summary graph. (2) **Meeting Summarization** [22] leverages agenda-based topics to segment meeting transcripts, and takes advantage of multi-modal sensing of the meeting environment, such as cameras to capture each participant's head pose and eye gaze. (3) **Multimedia News Question Answering** [23] employs multimedia event graphs to condition synthetic question-answer generation, and to automatically augment data via weak supervision.

5.1 WHAT IS TIMELINE SUMMARIZATION?

Timeline summarization [294, 296, 299, 300, 301, 302, 303, 304, 305, 306] aims at generating a sequence of major news events with their key dates from a large collection of related news from multiple perspectives (see Figure 5.5 for an example). The timeline summarization task poses several challenges to existing Natural Language Processing (NLP) techniques:

(1) In contrast to multi-document summarization (MDS) dealing with tens of documents [311], it summarizes hundreds of long documents, which requires the model to efficiently maintain a joint representation of the entire news collection, so that the summary has its coverage and coherence optimized globally.

(2) The summary is expected to select key dates and capture the temporal interdependency across key stories, which, compared to standard MDS, poses additional challenges in reconstructing temporal order.

(3) Manual labeling of timeline summaries is costly; thus the labeled data for model

training is very limited.

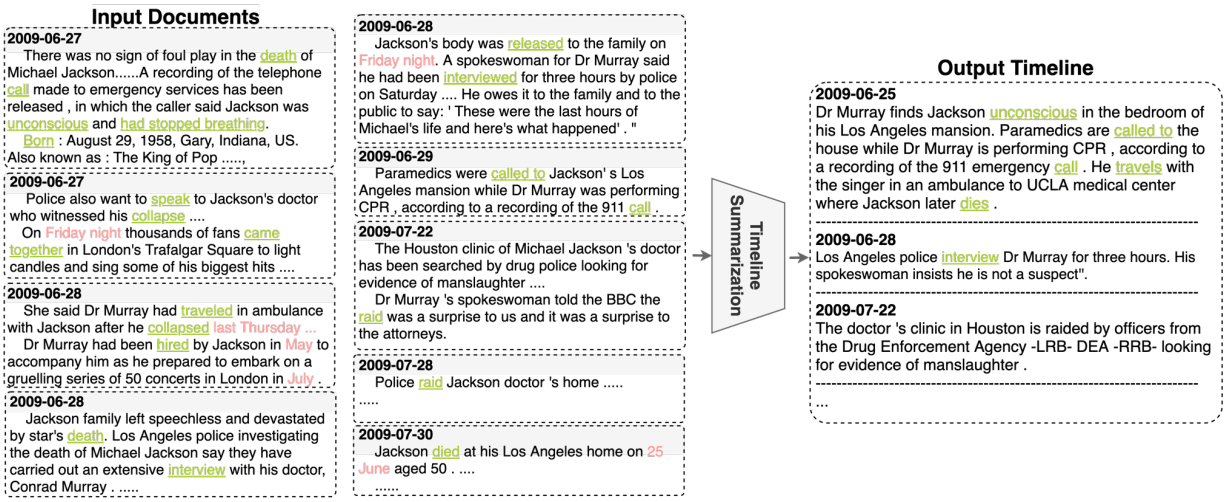


Figure 5.1: The input and output of timeline summarization.

	Multi-Document Summarization (MDS)	**Timeline Summarization (TLS)**
Size of Input Data	Tens of documents	Hundreds or thousands of long documents
Training Data	Available	No training set (due to annotation cost)
Temporal Dimension	None	- Select key dates - Capture the temporal interdependency across key stories
Knowledge Elements	Entities, relations, and events	Events are of higher priority

Table 5.1: Comparison between Timeline Summarization (TLS) and Multi-Document Summarization (MDS).

As a result, previous studies [305, 306] usually take an unsupervised approach. Specifically, these methods first identify the key dates from the publication time distribution. Then for each key date and its associated news articles, a summary is generated based on the salient sentences measured by the inter-similarity of these articles. In these methods, the document representations are limited to local text features, ignoring the global context of the news collection. The applications of neural models, especially advanced pre-trained language models, such as BERT [129], GPT-2 [346], GPT-3 [347], GPT-4 [12] are restricted

in terms of both long-distance representation capacity and memory efficiency when handling the global context within such input document size.

5.2 AN OVERVIEW OF EVENT GRAPH BASED TIMELINE SUMMARIZATION

We propose an event graph representation along with compression to deal with the representation difficulties in global graph contextualization, scalability, and time-awareness. Our solution consists of the following key ideas.

5.2.1 Event Graph Construction for Multi-doc Encoding

With state-of-the-art Information Extraction (IE) systems [162], we construct a single event graph from the input documents, with co-referential entities (e.g., *house*, *mansion* in Figure 5.2) and co-referential events (e.g., *die*, *collapsed*) merged across documents. Our comprehensive event graph connects events through temporal order (e.g., *interview* $\xrightarrow{\text{BEFORE}}$ *raid*), shared arguments (e.g., *called_to* $\xrightarrow{\text{AGENT}}$ *paramedics* $\xleftarrow{\text{PARTICIPANT}}$ *call*), and related arguments (e.g., *travel* $\xrightarrow{\text{DESTINATION}}$ *hospital* $\xrightarrow{\text{LOCATED\_IN}}$ *Los Angeles* $\xleftarrow{\text{AFFILIATION}}$ *police* $\xleftarrow{\text{PARTICIPANT}}$ *interview*). The graph structure enables the model to capture global long-distance interdependency between events across documents.

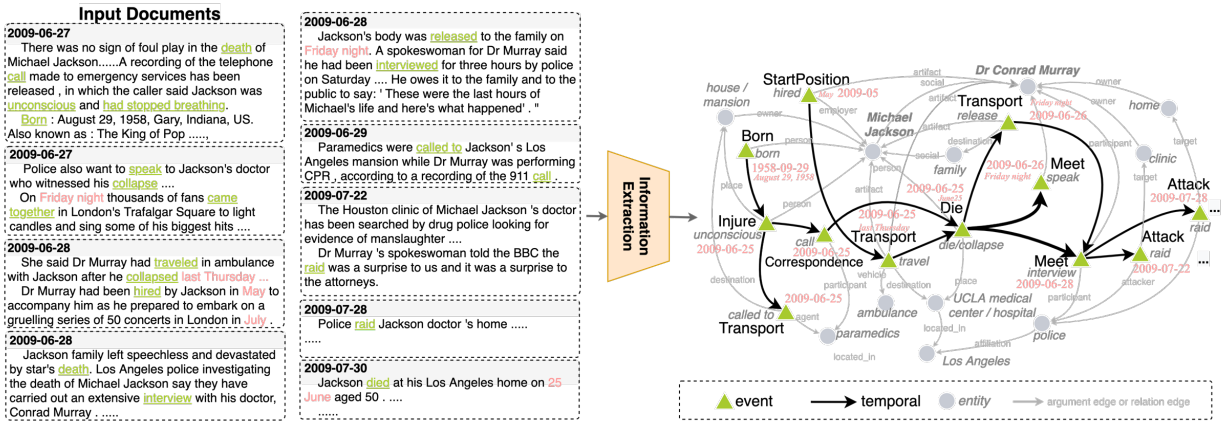


Figure 5.2: Event graph construction for multi-doc encoding.

5.2.2 Unsupervised event graph compression with optimal transport (OT)

We propose a new formulation of timeline summarization, by selecting event nodes from the input graph to form a smaller summary graph. Under a certain summary size constraint,

a summary graph with high coverage has a small information loss, compared to the one with low coverage [348]. We constrain the total number of event nodes to be kept in the summary, and optimize the summary graph to be close to the original graph using optimal transport. The training objective is to find the optimal transport plan between input and summary graph that has the minimal transport distance. Figure 5.3 shows an example of transporting node pairs in the input graph to the node pair $\langle die, interview \rangle$ in the summary graph. $\langle die, interview \rangle$ receives relatively large mass during the graph transport since it has small distance with multiple node pairs in the input graph, such as $\langle die, speak \rangle$. To obtain the minimal distance with only m events to be kept, a global decision is learned to select salient but also diverse events. The summary graphs are generated using a differentiable compression model according to a hyperparameter of compression rate, instead of using annotated timelines. Thus, our objective allows model training in an end-to-end unsupervised way.

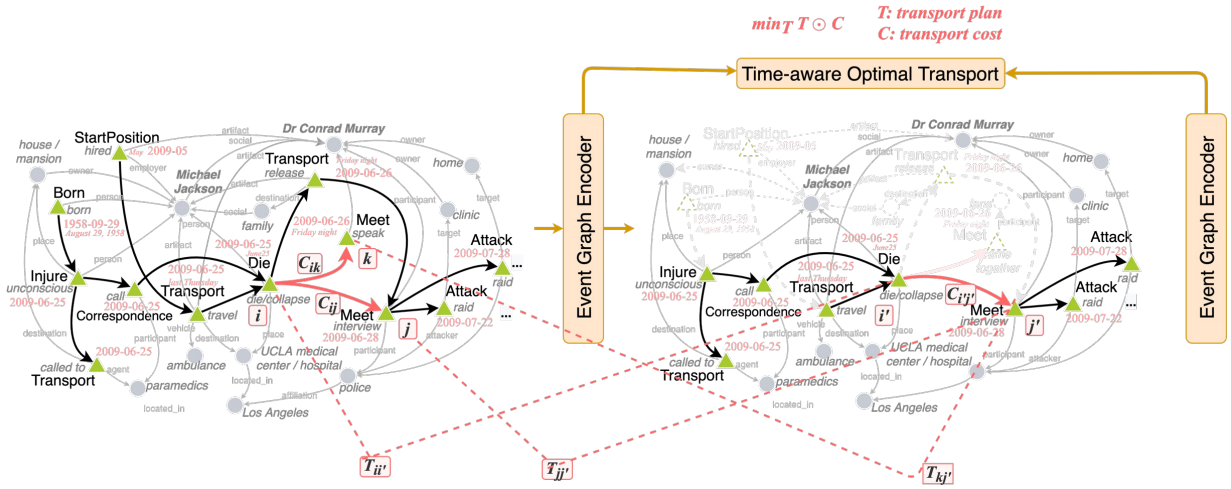


Figure 5.3: Event graph compression via optimal transport (OT).

5.2.3 Time-aware Gromov-Wasserstein Distance

The distance between two graphs should capture the following criteria:

i) Semantic relevance: Each node first has its initial *local* context encoded via a pre-trained BERT model and node type embeddings. For example, StartPosition event is not closely related to the Transport event in Figure 5.3 though they have temporal dependencies.

ii) Structural centrality: We employ a graph neural network to maintain a *global* context embedding by encoding the global structure topology, which enables the events of

high node centrality to gather comprehensive information from neighbors. For example, although both are MEET events, *interviewed* (by police) is more structurally salient than *speak*. It encodes the information not only from its neighbor events such as *raid*, but also from long-distance neighbors such as *travel* (to hospital) via the aforementioned argument paths.

iii) Temporal coherence: We define *time-aware Gromov-Wasserstein distance* over the temporal edges, and introduce a *temporal regularizer* to enlarge the distance between events that have wide time gap, such as the BORN and INJURE events in Figure 5.3, so that the temporal coherence can be captured. It enables the model to select temporally salient events that have temporal dependencies with multiple events in the news collection. Also, timeline summarization is sensitive to temporal ordering, such that the TRANSPORT (*traveling* in ambulance) before DIE in Figure 5.3 is more important to the story than the TRANSPORT (*releasing* body) after DIE. Hence, we distinguish the before and after events in the distance computation.

5.3 EVENT GRAPH CONSTRUCTION

The event graph is a heterogeneous graph G, where nodes are events $\{v_i\}$ and entities $\{e_j\}$, and edges contain event-event temporal ordering edges $\{\langle v_i, v_l \rangle\}$, event-entity argument edges $\{\langle v_i, a, e_j \rangle\}$, and entity-entity relation edges $\{\langle e_j, r, e_k \rangle\}$.

We apply OneIE [162], a state-of-the-art Information Extraction (IE) system, to extract entities, relations and events; then perform cross-document entity and event coreference resolution [327, 349, 350] over the document cluster of each timeline topic. We apply [351] to extract temporal relations for events in the same paragraph or having shared arguments. For example, *clashes* happen before *wound* given the sentence *fifty wounded are reported in the clashes*. To obtain the date of each event, We extract and normalize time expressions using publication date [335], and then apply [352] to extract the event temporal attributes from the context. If the temporal attributes can not be decided according to the context, we propagate the temporal attributes from neighbor events based on their shared arguments [352]. After that, we use the document publication date to populate the remaining missing dates. For example, in Figure 5.5, the date 2009-06-25 of the *collapse* (DIE) event is extracted from context *last Thursday*, and the date of the *unconscious* (INJURE) event is propagated along with their shared argument *Michael Jackson*.

5.4 EVENT GRAPH COMPRESSION BASED ON TIME-AWARE OPTIMAL TRANSPORT

5.4.1 Time-Aware Optimal Transport (OT)

Optimal Transport We aim to generate the summary graph S that has minimal OT distance with the input graph G, such that

$$D(G, S) = \min_{\boldsymbol{T}} \boldsymbol{T} \odot \boldsymbol{C}, \tag{5.1}$$

where $\odot$ represents the Hadamard product. $\boldsymbol{T} \in \mathbb{R}_{+}^{n \times m}$ denotes the transport plan, learned to optimize a *soft* node alignment between two graphs. Namely, each node in G can be transferred to multiple nodes in S with different weights. We use $T_{ii'}$ to denote the amount of mass shifted from node i in the input graph G to node i' in the summary graph S, as shown in Figure 5.5. $\boldsymbol{C} \in \mathbb{R}^{n \times m}$ is the cost matrix of event nodes between two graphs.

Time-Aware OT Distance Considering that event graphs are heterogeneous graphs, and timeline summarization is sensitive to temporal dependencies between events, we define the Gromov-Wasserstein Distance [339] on temporal edges to calculate distance between pairs of nodes within two graphs, i.e., $\langle i, j \rangle$ in G and $\langle i', j' \rangle$ in S:

$$D(G, S) = \min_{\boldsymbol{T}} \sum_{i,j \in G} \sum_{i',j' \in S} T_{ii'} T_{jj'} |C_{ij} - C_{i'j'}|. \tag{5.2}$$

Figure 5.5 shows an example of transporting edges $\langle i, j \rangle$ in the input graph to $\langle i', j' \rangle$ in the summary graph. The cost $|C_{ij} - C_{i'j'}|$ evaluates the intra-graph structural similarity between two pairs of nodes $\langle i, j \rangle$ in G and $\langle i', j' \rangle$ in S. To capture the direction of temporal ordering, we parameterize different matrices to distinguish the *before* and *after* nodes:

$$C_{ij} = \| \boldsymbol{W}_{\text{bfr}} \boldsymbol{v}_i - \boldsymbol{W}_{\text{aft}} \boldsymbol{v}_j \|_2 - \Omega(t_i, t_j). \tag{5.3}$$

In this way, although *travel* in Figure 5.5 and *release* are both TRANSPORT events connecting with the DIE event, they are distinguished during distance calculation. Here, $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are the node representations and we want them to capture the semantic relevance, structural salience and temporal coherence. As a result, we design an event graph encoder later in §5.4.2 from these three aspects.

Temporal Regularizer The OT distance between events should also capture temporal coherence. For example, in Figure 5.5, BORN event and INJURE event have large time gap, so that there should be a large distance between them, although they have direct connections in the graph. As a result, we use a regularizer $\Omega(t_i, t_j)$ to penalize events that have a large time difference $t_i - t_j$:

$$\Omega(t_i, t_j) = \frac{\beta}{(t_i - t_j)^2 + 1}, \tag{5.4}$$

where $\beta \in (0, 1]$ is a hyper-parameter.

5.4.2 Event Graph Encoder

In order to calculate the time-aware optimal transport distance, we encode both the input event graph and the summary graph to obtain the node representations, which capture text semantics, graph structures and preserves the temporal information.

Semantics Encoding To capture the local text semantics of an entity $\boldsymbol{e}$ or an event $\boldsymbol{v}$, we apply the pre-trained BERT [353] to initialize a contextualized embedding $\boldsymbol{w}$ using its text mentions. We use the average representation for nodes having multiple mentions, and concatenate it with the node type embedding $\boldsymbol{\phi}$, which is initialized by BERT using the type name. The frequency of events has been proven effective and critical to timeline summarization [305]. As a result, we add the number of its text mentions $|w|$ to capture the event frequency in the news collection:

$$\boldsymbol{v} = [\boldsymbol{w}_v; \boldsymbol{\phi}_v; |w_v|], \boldsymbol{e} = [\boldsymbol{w}_e; \boldsymbol{\phi}_e; |w_e|], \tag{5.5}$$

where $[;]$ denotes concatenation operation.

Graph Encoding After that, we employ an edge-wise graph neural network to contextualize all the nodes with their global graph contexts. We first generate edge type representation $\boldsymbol{a}$ and $\boldsymbol{r}$ by encoding the edge type name using pre-trained BERT, and temporal edge representation $\boldsymbol{t}$ is encoded using name "before". The message passed through an argument edge $\langle v_i, r, e_j \rangle$ is:

$$\boldsymbol{m}_{i,j} = \text{ReLU}\left[(\boldsymbol{W}_a\left[(\boldsymbol{v}_i - \boldsymbol{e}_j); \boldsymbol{a}\right]\right). \tag{5.6}$$

The messages of relation and temporal edges are similar, by replacing $\boldsymbol{a}$ with $\boldsymbol{r}$ and $\boldsymbol{t}$. We aggregate the messages using edge-aware attention following [267],

$$\alpha_{i,j} = \sigma(\text{MLP}(\boldsymbol{v}_i - \boldsymbol{v_j})), \tag{5.7}$$

where σ denotes sigmoid function. We adopt a two-layer MLP with ReLU as activation funtion.

The event node representation $\boldsymbol{v}_i$ is then updated using the messages from its local neighbors $N(v_i)$:

$$\boldsymbol{v}_i \leftarrow \text{GRU}\left(\left[\boldsymbol{v}_i; \sum\nolimits_{j \in N(v_i)} \alpha_{i,j} \boldsymbol{m}_{i,j}\right]\right), \tag{5.8}$$

similar to entity node representations.

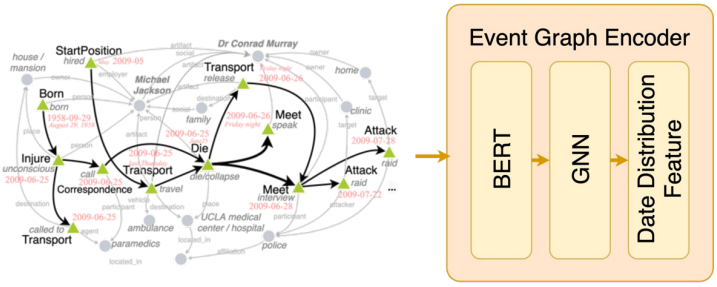


Figure 5.4: Graph encoding for event graphs.

Date Distribution Encoding To encode the date distribution, for each event v_i with date t_i, we concatenate the above node representation $\boldsymbol{v}_i$ with the number of documents published on t_i, the number of events happening on t_i, and the number of event text mentions attached to t_i in local context. It enables the OT distance to capture the corpus-level date salience.

5.4.3 Differentiable Graph Compression

To get a summary graph with m event nodes [16], we apply an event graph compression matrix $\boldsymbol{M} \in \mathbb{R}^{n \times m}$ following [354],

$$\boldsymbol{A}_S = \boldsymbol{M}^T \boldsymbol{A}_G \boldsymbol{M}, \tag{5.9}$$

[16] We only compress the event nodes since that the key for timeline summarization is salient event selection, while arguments are used to capture the distance between events.

where $\boldsymbol{A}_G \in \mathbb{R}^{n\times n}$ is the temporal edge adjacency matrix of event nodes in G, with $\boldsymbol{A}_S \in \mathbb{R}^{m\times m}$ for S similarly. For timeline summarization task, the parametrization of $\boldsymbol{M}$ has two requirements: (1) $\boldsymbol{M}$ is differentiable to enable end-to-end training; (2) we want to guarantee that the nodes in the summary graph are originally from the input graph (due to our extractive summarization goal), so we follow [354] to directly select nodes as summary nodes according to their weights $\boldsymbol{\alpha} \in \mathbb{R}^{n\times 1}$:

$$\boldsymbol{\alpha} = \sigma(\widehat{\boldsymbol{A}}\boldsymbol{V}\boldsymbol{W}_\alpha) \tag{5.10}$$

Here, $\widehat{\boldsymbol{A}} \in \mathbb{R}^{n\times n}$ is the normalized graph adjacency matrix defined in graph convolutional networks [355], $\boldsymbol{V} \in \mathbb{R}^{n\times d}$ is the node feature matrix, and $\boldsymbol{W}_\alpha \in \mathbb{R}^{d\times 1}$ is a parameter vector. σ is the sigmoid function.

We pick the top m values of $\boldsymbol{\alpha}$ and list them in the sorted order, denoted by $\boldsymbol{\alpha}_{sort} \in \mathbb{R}^{m\times 1}$. Similarly, $\widehat{\boldsymbol{A}}_{sort} \in \mathbb{R}^{n\times m}$ is the column-sorted and picked version of $\widehat{\boldsymbol{A}}$. Then the compression matrix $\boldsymbol{M}$ can be finally defined as

$$\boldsymbol{M} = \ell_1\text{-row-normalize}[\widehat{\boldsymbol{A}}_{sort} \odot (\mathbf{1}\boldsymbol{\alpha}_{sort}^T)], \tag{5.11}$$

where $\mathbf{1}$ means a column vector of all ones.

5.4.4 Training Objective

The optimal $\boldsymbol{T}$ that solves $D(G,S) = \min_{\boldsymbol{T}} \boldsymbol{T}\odot\boldsymbol{C}$ can be approximated by a differentiable Sinkhorn-Knopp algorithm [337, 338] following [339, 354],

$$\boldsymbol{T} = \text{diag}(\boldsymbol{p})\ \exp(-\boldsymbol{C}/\gamma)\ \text{diag}(\boldsymbol{q}), \tag{5.12}$$

where $\boldsymbol{p} \in \mathbb{R}_+^{n\times 1}$ and $\boldsymbol{q} \in \mathbb{R}_+^{m\times 1}$. The solution $\boldsymbol{T}$ can be computationally obtained by using Sinkhorn's algorithm. Starting with any positive vector $\boldsymbol{q}^0$ to perform the following iteration:

$$\begin{aligned} &\text{for}\ \ i = 0, 1, 2, \ldots \text{until convergence,} \\ &\quad \boldsymbol{p}^{i+1} = \mathbf{1} \oslash (\boldsymbol{K}\boldsymbol{q}^i), \\ &\quad \boldsymbol{q}^{i+1} = \mathbf{1} \oslash (\boldsymbol{K}^\top \boldsymbol{p}^{i+1}), \end{aligned} \tag{5.13}$$

where $\oslash$ denotes element-wise division. A computational $\boldsymbol{T}^k$ can be obtained by iterating a finite number k times,

$$\boldsymbol{T}^k := \text{diag}(\boldsymbol{p}^k)\boldsymbol{K}\text{diag}(\boldsymbol{q}^k). \tag{5.14}$$

The parameterization of the graph compression step and Sinkhorn-Knopp algorithm are differentiable, so we can optimize our time-aware optimal transport distance between two graphs in an end-to-end manner.

The advantage of our approach is that the training process is unsupervised, since the summary graph is generated automatically under the constraint of the hyperparameter m, i.e., the number of event nodes in the summary graph. The model parameters include those for the *graph encoder* (capturing semantic relevance, structural centrality and time salience), the *transport distance matrix* (capturing temporal coherence), the *compression model* (selecting top ranked nodes in a differentiable manner), and the *transport plan* (making a global decision to obtain minimum distance). They are optimized jointly to minimize the distance between the generated graph and the input graph.

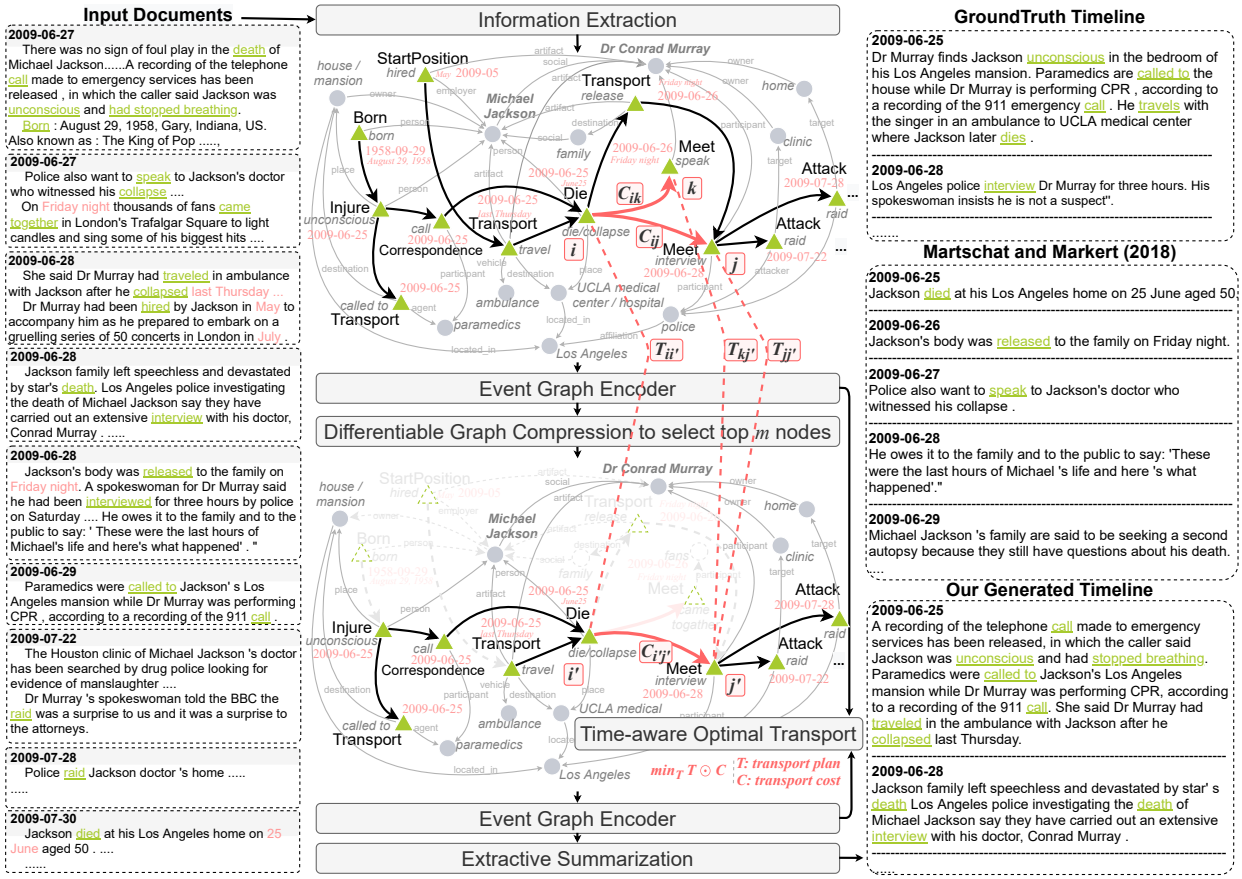


Figure 5.5: An example output of timeline summarization based on event graph compression.

5.4.5 Extractive Summarization

During summarization, the event summary graph is generated by selecting m events according to the event weights $\boldsymbol{\alpha}$, where m is a hyperparameter decided by the expected compression rate. To maintain the diversity of the temporal dimension following [305], we set a maximum event constraint to select no more than k events for each date. In detail, if the event number of one date reaches the limitation, the remaining events of that date will be ignored in the ranking list $\boldsymbol{\alpha}$, and only events happening on other dates can be selected to the summary graph. For each date, k is decided by the date distribution (i.e., the number of events happening on each date), as well as the compression rate hyperparameter.

Finally, for each event $v \in V_S$ in the summary graph, we extract an event summary sentence, i.e., the source sentence with the maximum event coverage.[17] The event summaries are ordered by dates to form the timeline. The event summaries on the same date are merged following the events' temporal orders with topological sort [356].

5.5 EXPERIMENTS

5.5.1 Experimental Settings

Datasets The evaluation is conducted on three datasets. *Timeline$_{17}$* [302] and *Crisis* [303] are two widely used timeline summarization datasets. *Timeline$_{17}$* contains 17 topics, and each topic has 1-3 ground-truth timelines, resulting in 19 timelines in total. *Crisis* has 5 topics and each topic has 4-7 ground-truth timelines, with 22 timelines annotated in total. We use all 19 and 22 timelines as references, and calculate the average scores following previous work.

To explore the robustness of our event graph compression for different scenarios, we also collect a new larger dataset *Timeline$_{100}$* containing 100 timelines from news websites including VoA[18] and Reuters[19]. The timelines are written by journalists and are manually curated. The dataset covers various topics related to the *economy*, *military*, *education*, etc. The input documents for each timeline are selected using BM25 [357]. For each dataset, we construct

[17] We select the events with highest temporal attribute accuracy if there is a tie. The events with temporal attributes extracted directly from the context are of highest priority, followed by events having temporal attributes propagated from neighbor events in Section 5.3, and then the ones using document publication date.

[18] `https://wwconw.voanews.com`

[19] `https://www.reuters.com`

input event graphs following Section 5.3. [20] We use the ACE event ontology[21], with 7 entity types, 6 relation types, 33 event types, and 22 argument roles. For the (unsupervised) training of our event graph compression model, we use event graphs constructed from VoA news between 2011 and 2017 [358]. The statistics are shown in Table 5.2.

Dataset	Split	#Document	#Event	#Entity	#Relation
Timeline_{17}	Input	4,650	74,320	115,585	136,509
	Timeline	19	974	1,936	1,134
Crisis	Input	20,463	325,695	551,228	610,410
	Timeline	22	736	1,184	1,309
Timeline_{100}	Input	10,379	178,581	301,132	306,975
	Timeline	100	3,296	8,901	23,732
Unlabeled (for OT)	Input	72,576	913,679	381,735	1,046,066
	Timeline	-	-	-	-

Table 5.2: Data statistics of timeline summarization benchmarks.

Evaluation Metrics We use the conventional metrics for timeline summarization [305] to evaluate the key date selection using *Date* F_1 and the content generation using ROUGE scores, including (1) *concat* F_1 to compute ROUGE by concatenating the summaries of all selected dates; (2) *agree* F_1 to compute ROUGE only between the summaries which have the same dates; (3) *align* F_1 to first align summaries in the output with those in the reference based on similarity and the distance between their dates, then compute the ROUGE score between aligned summaries. Distant alignments are punished.

Baselines We include the following baselines:

(1) Chieu [299], a typical extractive model based on sentence similarity;

(2) Submodular [305], the state-of-the-art extractive timeline sumarization model based on submodular functions;

(3) PacSum [359], the state-of-the-art unsupervised graph-based ranking summarization baseline, which utilizes BERT to encode sentences for sentence centrality ranking in a sentence graph. We use the publication date of the selected sentence as key dates;

(4) SummPip [360], the state-of-the-art unsupervised multi-document summarization baseline, which constructs a sentence graph and performs spectral clustering. After that, a sum-

[20] The preprocessed event graphs are released together with the dataset.

[21] `https://www.ldc.upenn.edu/collaborations/past-projects/ace`

mary is generated for each sentence cluster by multi-sentence compression, and we use the most frequent publication date of the sentences in the cluster as key dates;

(5) "w/o temporal regularizer", an ablation study by removing the temporal regularizer in the OT distance. [22]

Training Details The dimension of contextual embedding, type embedding, and edge embedding are 768. β is 0.5. γ is 1. The ratio of event nodes kept after compression m is determined based on the ratio of input graph size and summary graph size of the dataset. We use 0.05 for *Timeline*$_{17}$ dataset, 0.005 for *Crisis* dataset, and 0.05 for *Timeline*$_{100}$ dataset [23]. Due to the large size of input graphs, we first compress the subgraph extracted from each publication date following the hard cutoff of [305], and then compress the graph of the entire corpus. The graph compression model is trained on one Tesla V100 GPU with 16GB DRAM.

5.5.2 Quantitative Performance

As shown in Table 5.3, our method outperforms baselines on all three datasets. Event graph connects events through entities and temporal relations, which enables capturing the correspondence between events, and excludes unrelated events. General multi-document summarization and text graph based summarization cannot capture the temporal dimension, so the performance is especially low on date F_1, agree F_1 and align F_1. All Concat F_1 scores are significantly different from baselines with p value less than 0.05.

Removing the temporal regularizer results in a consistent performance drop on date F1, showing that our time-aware OT helps select events that are temporally coherent.

We achieve larger gains compared to baselines on *Crisis* dataset, which has larger input graph size and compression rate according to Table 5.2. It proves the effectiveness of our event graph on encoding a large number of documents and perform effective summarization. Compared to Timeline$_{17}$, the performance gain on Timeline$_{100}$ is larger, which cover more scenarios. It demonstrates the robustness of our event graph compression method.

[22] For fair comparison, our baselines focus on unsupervised methods that can produce key dates, which excludes text word graph based models and pretrained langauage model based generation models due to lack of temporal dimensions.

[23] We choose m based on three times of reference compression rates to allow comprehensive information being kept.

Dataset	Model	Concat F_1 R-1	Concat F_1 R-2	Agree F_1 R-1	Agree F_1 R-2	Align F_1 R-1	Align F_1 R-2	Date F_1
Timeline_{17}	Chieu [299]	0.223	0.049	0.024	0.008	0.046	0.012	0.195
	Submodular [305]	0.364	0.087	0.092	0.021	0.103	0.024	0.543
	PacSum [359]	0.231	0.054	0.029	0.012	0.035	0.013	0.173
	SummPip [360]	0.242	0.057	0.028	0.009	0.030	0.007	0.158
	Optimal Transport	0.370	0.089	0.092	0.020	0.103	0.024	0.550
	w/o temporal regularizer	0.369	0.087	0.091	0.018	0.101	0.025	0.545
Crisis	Chieu [299]	0.348	0.065	0.026	0.006	0.047	0.010	0.146
	Submodular [305]	0.333	0.071	0.056	0.012	0.076	0.015	0.288
	PacSum [359]	0.144	0.017	0.004	0.001	0.008	0.001	0.077
	Summpip [360]	0.124	0.016	0.004	0.001	0.007	0.001	0.069
	Optimal Transport	0.348	0.074	0.058	0.012	0.079	0.015	0.291
	w/o temporal regularizer	0.348	0.073	0.056	0.011	0.076	0.014	0.290
Timeline_{100}	Chieu [299]	0.127	0.028	0.011	0.003	0.017	0.004	0.138
	Submodular [305]	0.257	0.060	0.016	0.005	0.021	0.007	0.290
	PacSum [359]	0.219	0.045	0.011	0.002	0.016	0.005	0.151
	SummPip [360]	0.196	0.034	0.011	0.002	0.017	0.004	0.158
	Optimal Transport	0.278	0.067	0.017	0.005	0.023	0.008	0.295
	w/o temporal regularizer	0.279	0.067	0.015	0.004	0.021	0.007	0.292

Table 5.3: Performance on timeline summarization.

5.5.3 Qualitative Analysis

Compared with the baseline timelines shown in Table 5.4, the date selection and event node coverage of our method are much higher compared to baselines. The event triggers are highlighted in red for easier comparison. Figure 5.5 show the generated timeline of our walk-through example in this chapter, with the comparison with the reference timeline and the best performing baseline [305]. The number of dates selected by the baseline is larger compared to our approach, which demonstrates that our approach can better detect salience of dates. We think this is because we take advantage of event graphs to capture the events that are temporally salient. For example, our approach avoids the dates that do not have associated salient events, such as 2009-06-26. Also, our temporal attributes are more comprehensive and accurate due to the attribute propagation through shared arguments. For example, the dates of *unconscious* and *travel* in Figure 5.5 are propagated from the *die* event via the shared argument *Michael Jackson*.

Compared to the baselines, our approach keeps more events in the summary (highlighted in green in Figure 5.5), while the baseline may produce a summary without events included,

e.g., the summary of 2009-06-29.

Compared to the reference timeline, our model is shown to successfully detect the salient events in the graph compression process. Although the *release* event has connections to multiple events, it is not semantically relevant to other events, and thus it will not receive a large mass during the transportation. The *speak* event is not strongly connected to other nodes, and it is semantically close to *interview*, which will not be selected in the global decision of the optimal transport plan. Similarly, the *born* event is omitted due to its large time gap with other events, and the *hire* event is excluded since it is not semantically related to other events.

Method	**Example Output**
Reference	**2011-02-18** Libyan state television shows images of men chanting pro-Gadhafi slogans , waving flags and singing around the Libyan leader 's limousine as it creeps through Tripoli . In Benghazi , human rights groups and protesters claim they 're under attack by pro-government security forces . Among the tens of thousands of protesters who take to the streets , at least 20 people are killed and 200 are wounded , according to medical sources . **2011-02-19** Protests continue to turn violent , however the death and injury toll is unclear . In Benghazi , witnesses report bloody clashes with soldiers firing tear gas and bullets . Witnesses say protests have erupted in cities across the country . Human Rights Watch reports that 84 people have been killed in Libyan demonstrations since February 15 . **2011-02-20** Violence surges in Benghazi where a witness says protesters have taken control of the city and much of Tripoli . Gadhafi 's son Saif al-Islam Gadhafi appears on state television to warn demonstrators that the country could fall into civil war if their protests do not subside .

Continued on next page

Table 5.4: Output comparison between our system and baselines.

Method	Example Output
	2011-02-21 The Libyan newspaper Quryna reports that the country 's justice minister has resigned to protest what he calls a " bloody situation and use of excessive force " by security forces against protesters.
Chieu [299]	**2011-02-21** By the CNN Wire Staff Libya protests spread to Tripoli State Department has ordered the evacuation of all non-essential personnel The Obama administration is stressing the need to avoid violence against protesters Gadhafi 's son has warned of a possible civil war if protesters do not back down Washington (CNN) – The United States on Monday condemned the violence in Libya and called for a halt to the " unacceptable bloodshed " in response to civil unrest , Secretary of State Hillary Clinton said in a statement .
Submodular [305]	**2011-02-15** Protests began February 15 in the eastern city of Benghazi , Libya 's second largest . Witness says square in Benghazi is full of protesters , but there is little sign of police or military Tanks surrounded demonstrators in Benghazi , a protester says 50 reportedly killed since Tuesday , 20 of them Friday U.S. president condemns the government crackdowns in Libya , Bahrain and Yemen (CNN) – At least 20 people were killed and 200 more were injured Friday in the northern Mediterranean city of Benghazi , Libya 's second-largest , said a medical source in Benghazi who was not identified for security reasons . **2011-02-21** Among other things , Washington was taking a close look at a speech early Monday by Saif al-Islam Gadhafi – the Libyan leader 's son – which included warnings of a civil war if demonstrations in the North African country do n't stop . The United States on Monday condemned the violence in Libya and called for a halt to the " unacceptable bloodshed " in response to civil unrest , Secretary of State Hillary Clinton said in a statement .
Ours	**2011-02-16**

Table 5.4 (cont.)

Method	Example Output
	Source : Several people arrested after police confronted protesters in Benghazi , Libya . **2011-02-18** An Iranian opposition member warns that street protests could lead to civil war " Nastaran " warns that protests are strengthening Iran ' s Revolutionary Guard and pro - government militia. **2011-02-19** A Libyan woman supportive of the protesters , who was not identified to protect her safety , told CNN that army soldiers on Saturday initially claimed solidarity with the demonstrators , only to reverse their tack and open fire on the crowd . Three of those injured are in critical condition , the sources said . While Human Rights Watch , citing interviews with hospital staff and witnesses , reported 84 deaths since Tuesday , the total number is unknown and could n ' t be independently confirmed by CNN . **2011-02-20** Protests continue to turn violent , however the death and injury toll is unclear .

Table 5.4 (cont.)

5.5.4 Human Evaluation

We follow previous work [306] to do a scoring-based evaluation. We instruct the human annotators to read 15 randomly sampled reference timelines, and rate summaries generated by our system and baselines on a 1-5 point scale (1 is the worst and 5 is the best). We provide reference timelines as the gold standard to annotators, instead of providing the input news collection. It is because that each timeline contains hundreds of long documents as input, making it hard to judge coverage and control scoring standards of multiple annotators. As the evaluation is scoring-based, we only ask one annotator to score all timelines of each topic to guarantee the same scoring standard. The order of annotating timelines is random, and the annotators have no knowledge about the order of the systems. Each timeline annotation takes around thirty minutes.

The timelines are evaluated in the following dimensions: (1) *general score*: the general quality of the timeline; (2) *coverage score*: the events that are covered by the timeline; (3)

Model	General	Coverage	Coherence	Temporal Preserving
Chieu [299]	2.4	1.4	2.5	1.4
Submodular [305]	3.2	2.7	3.4	2.6
Optimal Transport	3.9	2.9	3.7	2.8

Table 5.5: Human evaluation on a scale of 1-5 (1 is the worst and 5 is the best).

coherence score: the coherence of the story; (4) *temporal preserving score*: the selection of key dates. Table 5.5 shows that our approach gets better results on all four measures, proving that our model is reasonable to find semantically relevant, structurally salient and temporally coherent events.

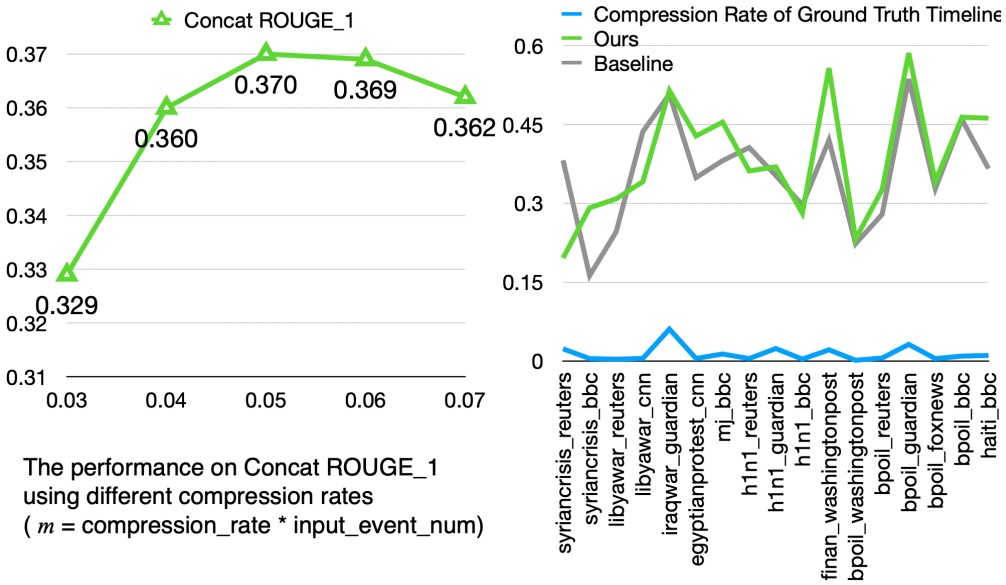


Figure 5.6: Analysis about compression rates.

5.5.5 Discussions

Generation Length. Previous work on timeline summarization [299, 305] relies on the reference timeline to decide the compression parameters, such as the overall length or the number of days. In our model, the number of nodes to be kept is decided by the hyperparameter m. Following previous work, we choose m based on the reference compression rate, i.e., the ratio of the event nodes in reference summary to the input event nodes, as detailed in Section 5.5.1. Figure 5.6 shows the relevance between the performance and compression rate.

Topic: BP Oil Spill	Graph Size	Concat ROUGE_1		
		Ours	Baseline	Δ
bpoil_washingtonpost	2582	0.232	0.223	+0.009
bpoil_guardian	2744	0.566	0.536	+0.029
bpoil_bbc	2972	0.464	0.459	+0.004
bpoil_foxnews	3032	0.341	0.328	+0.014
bpoil_reuters	3488	0.326	0.279	+0.047

Table 5.6: Analysis on the size of input event graph.

Compression Rate. The summarization performance is affected by the compression rate of the reference summary. Figure 5.6 shows that our model achieves larger gains compared to baselines on the timeline with higher reference compression rates, demonstrating that our model is able to effectively select salient events for a large input corpus.

Timeline Topics. Figure 5.6 shows that the compression rates do not have correlations with timeline topics, and our performance gains compared to baselines are not closely related to timeline topics, proving the robustness of our method.

IE Quality. We use state-of-the-art IE model [162] for the event graph construction. The IE quality on *Finance* is higher, leading to larger gains compared to baselines.

Input Graph Size. When generating timelines for the same complex event *BP Oil Spill*, as shown in Table 5.6, the performance gain is generally increasing with respect to the input graph size. It proves the effectiveness of our model on selecting salient information from large graphs.

5.6 CONCLUSIONS AND FUTURE WORK

In this chapter, we present a novel event graph compression framework for timeline summarization and achieve state-of-the-art on multiple real-world datasets. Our usage of event graphs allows for efficient joint encoding of a large number of documents; and our proposed time-aware optimal transport allows unsupervised training of the entire framework. It is the first study to use event graph representations to overcome fundamental challenges in handling massive unstructured data that exist in various applications. It provides tangible guidelines to use event structural knowledge in practice, and shows positive results on long-standing open problems in event tracking.

Our future work involves broadening our approach to include general fact-based generation and abstractive summarization. We also recognize the need to address the limitations of current large models. Although large models are expanding in terms of input length, they're

not necessarily improving their capacity to capture long-distance knowledge and its interconnections. This prompts the need to devise a strategy to comprehend how information is interconnected and to further discern which pieces of information hold more importance.

Moreover, the modeling of the temporal dimension needs further improvements, as it plays a significant role in understanding social context. It helps to trace the sequence of events, providing a clearer understanding of how one event leads to another. Therefore, we plan to integrate richer semantics into the edges of our models, ensuring that the relationships between events are not just sequential but also causally linked.

Additionally, we aim to incorporate subevent relationships for hierarchical timeline generation and include more multi-modal information in the source data. This comprehensive approach would not only enhance the coherence and structure of the timeline but also enrich our datasets and potentially improve the performance of our models. By considering these various dimensions, we hope to provide a more holistic and in-depth understanding of the information.

CHAPTER 6: CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

6.1 CONCLUSIONS

The knowledge of events is scattered in a variety of languages and data modalities. The limitation of computers as being more knowledgeable is not a lack of knowledge, but rather the inability to synthesize all scattered information in order to apply it effectively.

Traditionally, multimodal factual knowledge extraction has been entity-centric with a focus on concrete concepts (such as objects, object types, physical relations, e.g., a person in a car), but lacks ability to understand abstract semantics (such as events and semantic roles of objects, e.g., driver, passenger). However, such event-centric semantics are the core knowledge communicated, regardless whether in the form of text, images, videos, or other data modalities. However, existing methods oversimplify event understanding to be single-modal (text-only or vision-only), local, sequential and flat. However, real events are multimodal, hierarchical and probabilistic.

Hence, at the core of this thesis research in Multimodal Information Extraction (IE) is to bring such deep factual knowledge view to the multimodal world. Modeling event semantics in multimodal data poses significant challenges including:

- **Complexity**: The complexity stems from the involvement of multiple modalities. Since events are closely related to status changes, reading the complex situation requires to process information of both text and vision modalities and consolidate complex semantic structures across various modalities.

- **Dynamics**: One primary challenge of event-centric understanding is the dynamic nature of events, which requires to model the temporal dynamics spanning over a long horizon.

- **Factuality**: Fact-based event summarization is a major bottleneck for current large-scale pretraining models, specifically the capability to trace back the knowledge within the generated output to ensure its factual consistency with the input.

To address these issues, we propose three principles on modeling multimodal event semantics as the core of this thesis:

- **Zero-Shot Cross-Modal Structured Representation**: Our work endows machines to understand complex abstract semantic structures that are difficult to ground into

image regions but are essential knowledge (such as events and semantic roles of objects) It is able to consolidate complex semantic structures of multiple modalities, providing a major benefit over recent research advances in single-modality (text-only or vision-only) knowledge.

- **Temporal Event Graph Model as Schema**: We propose to learn a graph model to learn the process of event evolution along the timeline. It can input partially instantiated graphs to "grow" the graph either forward or backward in time to predict missing events, arguments, or relations, both from the past and in the future. Graph structure can capture transitions between events in a long horizon, and has better explanability. It empowers machines to perform probabilistic modeling of event prediction.

- **Fact-Based Event Summarization**: After converting unstructured data to structured events, we leverage the global event graphs to support downstream tasks, such as timeline summarization, meeting summarization, questions answering, etc.

Guided by these principles, in this thesis we study three fundamental and highly related event-centric tasks, event extraction, event schema induction, and event summarization. We proposed the following models to tackle the aforementioned problems in complexity, dynamics, and factuality:

(1) understanding multimodal semantic structures that are abstract (such as events and semantic roles of objects): we propose zero-shot cross-modal transfer (CLIP-Event), which is the first to model event semantic structures for vision-language pretraining, and supports zero-shot multimodal event extraction for the first time;

(2) understanding long-horizon temporal dynamics: we introduce Event Graph Model, which empowers machines to capture complex timelines, intertwined relations and multiple alternative outcomes. I have also shown the positive results of event-centric knowledge on long-standing open problems, such as timeline generation, meeting summarization, and question answering.

(3) generating fact-based summaries: We show its positive results on long-standing open problems, such as timeline generation, meeting summarization, and question answering.

In this way, our research enables machines to move from surface semantic understanding to deep semantic understanding. So that machines can read complex multimodal semantics, think long and wide with a event graph context, and write organized and grounded with knowledge tracing back to the source.

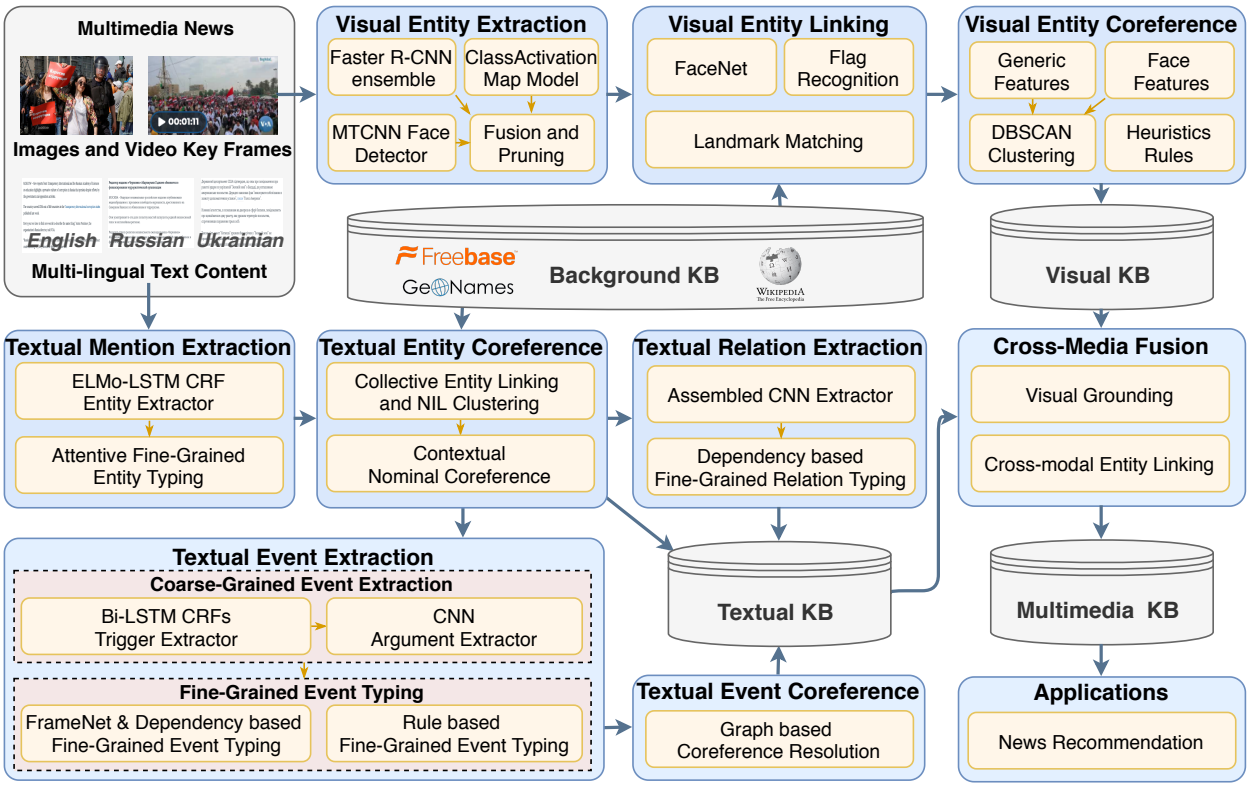


Figure 6.1: The model architecture of GAIA system.

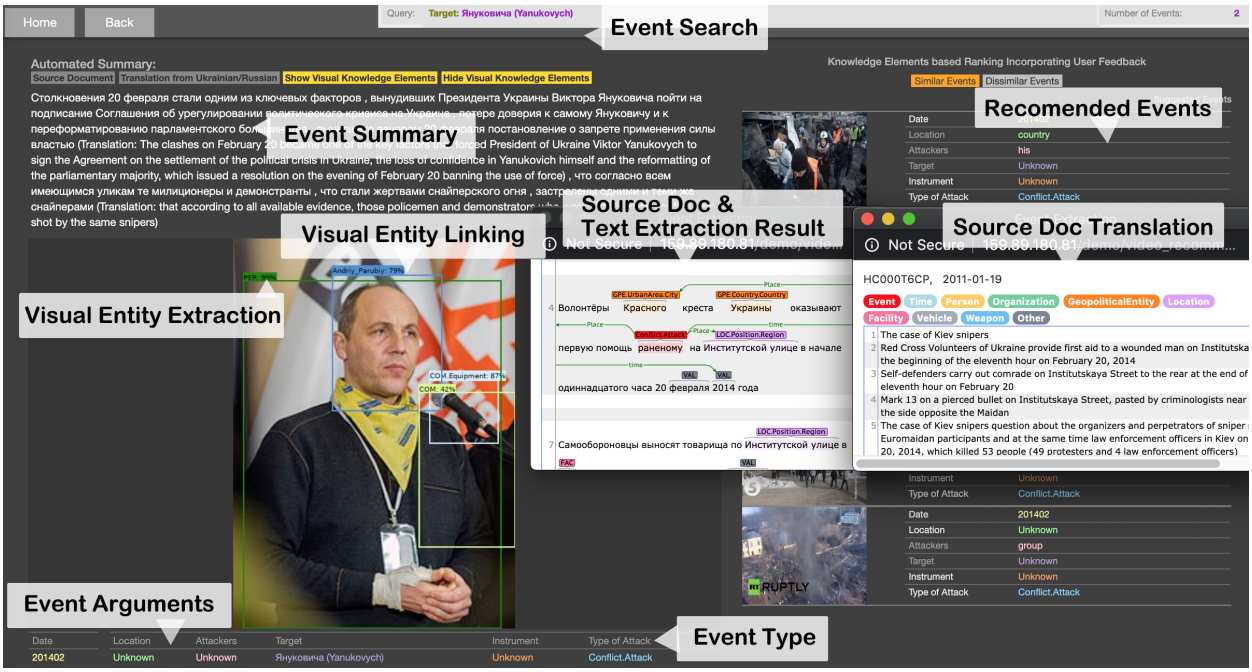


Figure 6.2: User-facing views of multimodal event knowledge extraction.

6.2 APPLICATIONS

Our models and algorithms for information extraction have been successfully applied to a serious of knowledge extraction systems and have won top performance at NIST SM-KBP evaluations of multiple years.

6.2.1 GAIA: Multimodal Knowledge Extraction System

GAIA is an open-source multimedia knowledge extraction system. Knowledge Extraction aims to discover structured knowledge elements such as entities, relations, and events from unstructured data, and link them to external knowledge bases such as Wikipedia, and GeoNames. Knowledge extraction outputs can support a wide range of downstream applicaions. In addition to the text IE models mentioned in this thesis, the GAIA system has also been inherently designed for multi-media. We extract complementary knowledge from texts, images, and video frames, and integrate the knowledge across modalities. Meanwhile, following a rich ontology, the GAIA system is able to extract fine-grained types which is crucial to scenario understanding and event prediction. As Figure 6.2 shows, GAIA contains two major modules: (1) Text Knowledge Extraction (entity extraction and coreference, relation extraction, and event extraction and coreference) and (2) Visual Knowledge Extraction (entity extraction, entity linking, and entity coreference). GAIA received the Best Demo Paper Award at ACL2020.

6.2.2 Event Tracking via Timeline and Heat Map

Generally, a single document can only provide limited information even with a perfect information extraction system. It is desirable to extract and aggregate knowledge elements across multiple languages and documents to build a complete view towards a specific event, scenario, or topic. Therefore, we develop a comprehensive multilingual knowledge extraction, aggregation, and visualization system as Figure 6.3 shows. The system performs entity discovery and linking, time expression extraction and normalization, relation extraction, event extraction, and event coreference. The system supports the extraction of 7 entity types, 23 relations, and 47 event types.

We extract and normalize time arguments to construct an event timeline in Figure 6.4 using TimelineJS for visualization.[24] There are three zones in the web-enabled timeline interface. By clicking on an event in the timeline, the pertinent context sentence for that

[24] `https://timeline.knightlab.com/`

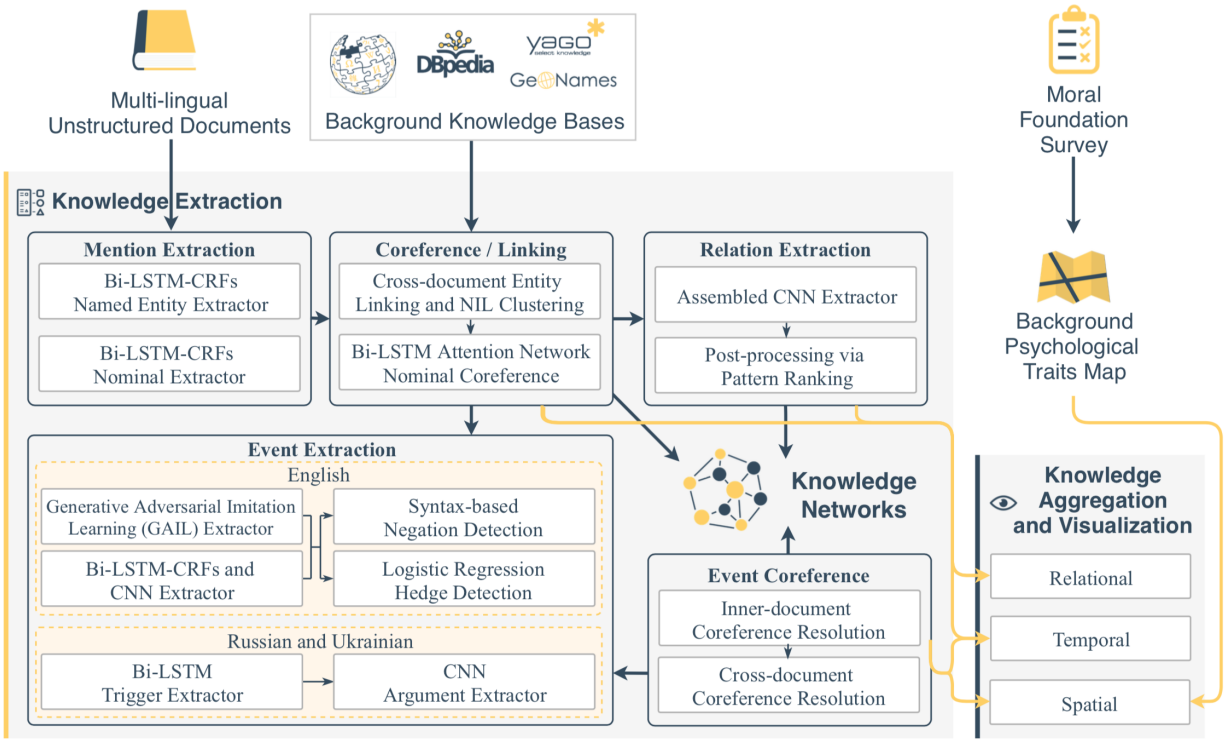


Figure 6.3: The model architecture of event tracking system.

event is displayed in the middle of the screen with the trigger and arguments highlighted in color, along with a link to the sentence's source document. Clicking on the source document link retrieves the document with full inline annotations and its publication date, to support inference of the absolute date(s) from relative time expressions in the text.

6.3 THE FUTURE OF EVENT-CENTRIC MULTIMODAL UNDERSTANDING

In this section, I present my vision for the future of event-centric multimodal information access, through a structured knowledge view that is easily explainable, highly compositional, and capable of long-horizon reasoning.

6.3.1 The Beauty of IE and Its Future

Information Extraction (IE) aims to uncover important facts that satisfy people's information needs, as well as connecting those facts in a way that provides people with meaningful hypotheses. The beauty of IE lies in its ability to provide a knowledge-intense structure, highlighting critical areas to focus on, especially when dealing with large corpora and long-

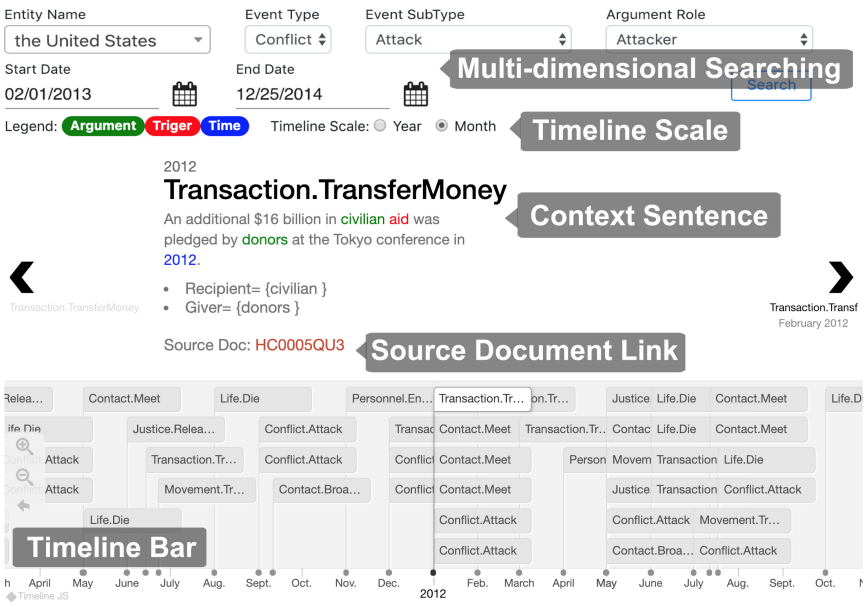


Figure 6.4: The timeline tracking interface.

distance captures. It is this capability that enables us to reach a comprehensive understanding of situations. IE successfully amalgamates knowledge from multiple languages and data modalities, fostering processes such as planning and reasoning. Various commercial assistant technologies, such as Siri, heavily rely on structured knowledge for their functionality.

However, currently available IE tools tend to be restrictive in their ontologies, lack general-purpose scalability, and are scarcely found in popular Natural Language Processing (NLP) packages like spacy and Stanza, where named entity recognition is a standard component. This creates a significant gap as IE tools are crucial for situation understanding and should be more readily accessible to users. One reason is that the strict structure has been one of the major obstacles in preventing IE from evolving into an adaptable off-the-shelf tool. This is because it attempts to precisely and discretely represent semantics across various levels of granularity, which is difficult even for humans to create annotations. The heavy reliance on ontology makes it difficult to generalize and to capture the most appropriate level of granularity.

So one future direction to extend IE to a more semi-structured representation. Each node can represent a natural sentence (allowing us to capture different levels of semantics) and each edge can represent any level of semantics (e.g. hierarchical, temporal, or causal edges via NLI models). With this approach, we are still able to keep the most beneficial aspects of IE, such as the links between long-distance concepts, while simultaneously capturing semantics of various granularities.

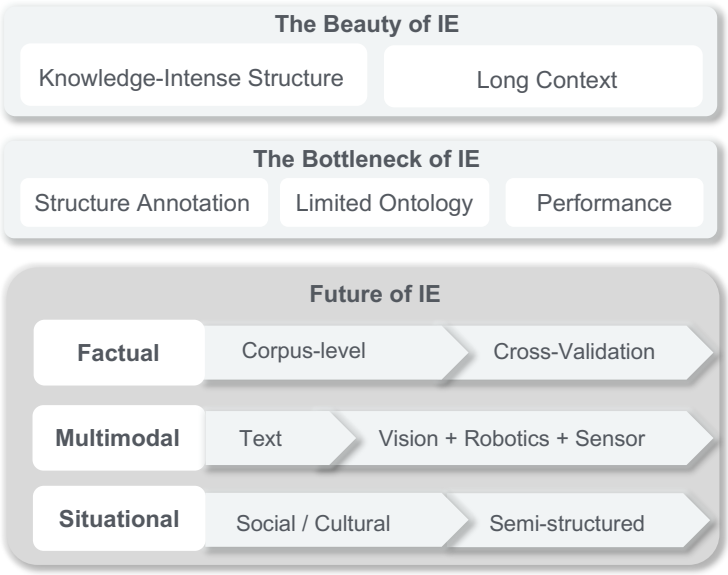


Figure 6.5: The future of Information Extraction.

Another improvement in need is the evaluation process. Traditional evaluation of thumbs up/down is too coarse-grained and not an effective way to use human intelligence. The goal of IE is to uncover important facts that satisfy people's information needs, so the evaluation process should focus less on annotating facts, but focus more on assessing whether machines are connecting and leveraging those facts in a way that provides people with meaningful hypotheses. Also, grounding evaluations to user scenarios will be important to evaluate whether models are really understanding and leveraging knowledge, rather than just capture the surface data distribution.

6.3.2 Factuality in Information Access

One major future research direction is to improve factuality of information access. The aforementioned semi-structured knowledge representation format can seamlessly integrate with large language models, offering a convenient way to generate factual content. Moreover, this representation enables human involvement in verifying and enhancing knowledge representation, fostering collaboration between humans and machines. Investigating how factual knowledge can be extracted and leveraged to promote truthfulness in generated content is an important research direction.

A primary solution of ensuring truthfulness is using both external and internal knowledge. As shown in Figure 6.6, knowledge can be presented from different sources. External knowledge includes informative context, knowledge bases, and related documents from open

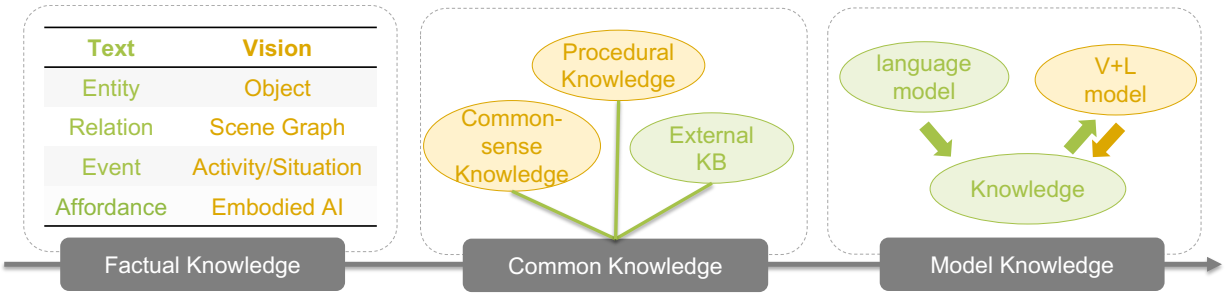


Figure 6.6: Knowledge categorization based on knowledge sources.

web resources, typically relying on the success of information retrieval [361, 362, 363, 364], information extraction [365], grounded generation [366, 367, 368] and knowledge-augmented generation [369, 370]. Internal knowledge involves the implicit parametric knowledge stored within the model, where the correction and refinement of parametric knowledge remains very limited [371, 372, 373, 374, 375], and the explorations are limited to the inference stage solely. Nevertheless, knowledge consists of more than just triples. It features complex structures, semantically rich edges, as well as causal/temporal indicators. To address this problem, it is crucial to not only decipher how intricate structured knowledge is interpreted through model parameter patterns, but also understand how the model pieces knowledge together and governs the underlying logic during generation. A significant challenge in knowledge-controlled generation is defining an appropriate knowledge representation that features both complex structures and distributed representations. This representation should combine the strength of symbolic-based reasoning to minimize unwarranted inferences, as well as the flexibility of distributed representations to encode any semantic granularity. Both these features have proven to be essential [376]. The ultimate goal is to rectify errors and refine model parameters by tracing back to the original pieces of knowledge, thereby enabling control over content generation to ensure factual accuracy and minimize hallucinations. These challenges are of utmost importance when utilizing large language models.

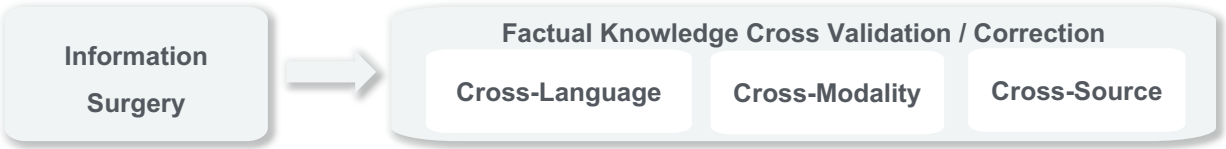


Figure 6.7: Factuality, Evidence and Truth in information access.

Another important aspect to improve factuality and truthfulness of information is through the verification of facts using multiple modalities, as shown in Figure 6.7. For example, speech plays a crucial role among these modalities, as it can effectively capture human

emotions and reactions that provide valuable signals for fact validation. There is a great opportunity to integrate multiple modalities to cross-validate facts across speech, video, and text to ensure their reliability and truthfulness.

6.3.3 Modeling Semantics of the Physical World

Semantic understanding and language acquisition in humans does not solely rely on reading; it also occurs through active engagement with the surrounding environment. In light of the remarkable progress of the language world, we are now able to transit towards complex and much more modalities that were previously beyond consideration. In this section, we aim to extend the scope of information access by incorporating other modality signals in the physical world. The ultimate research goal is to make machines capable of understanding deep semantics as humans do, especially abstract semantics in the real world, including semantic roles of objects (such as *victim*, *detainee*), and semantics of abstract concepts (such as *love*, *happiness*). With this structured view of knowledge, machines are able to further comprehend, reason, and communicate knowledge through vision and natural language.

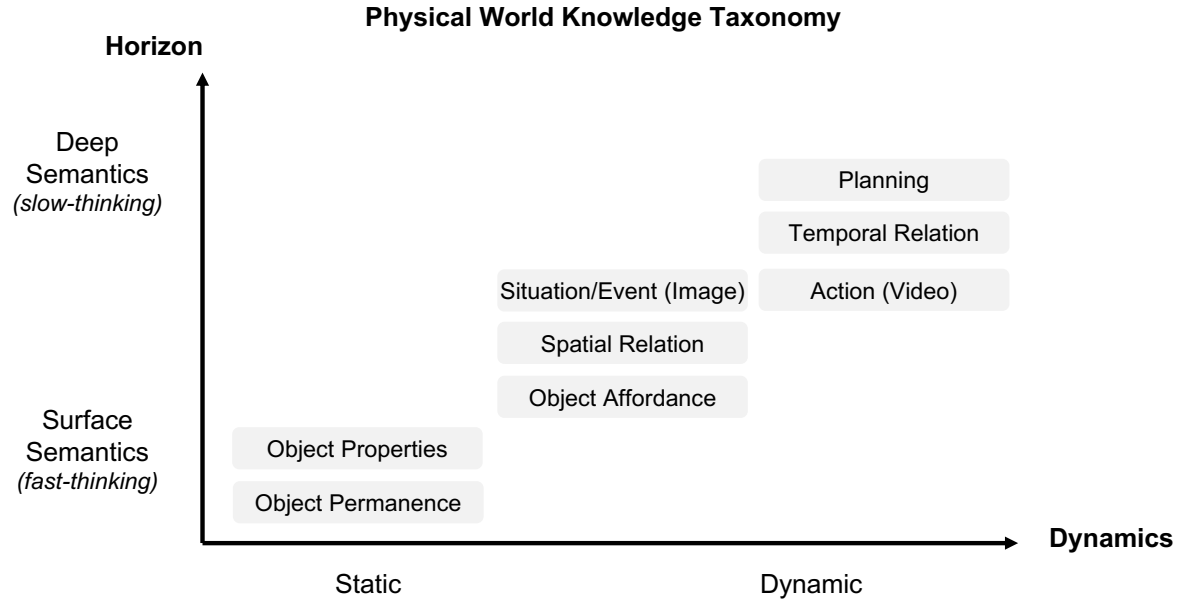


Figure 6.8: The knowledge to be acquired about the physical world.

Abstract Semantics Modeling: It is crucial to build open-world vision-language and reasoning models that reason about abstract concepts, from general, simple, and observational to specific, complex, and interpretive. It is desirable to develop a neural symbolic reasoning framework that is able to compositionally learn new concepts with the training

signal transferred from language. It will learn not only to discriminate between known concepts, but also to derive unnamed associations that provide the foundation for later learning of novel concepts. Reasoning will be performed under a graph structural context, capturing semantic roles, attribute semantics, temporal orders, and more. This approach can be regarded as utilizing *Multimodal Semantic Parsing* to model fine-grained cross-media relationships, including the sub-graph structures. There is a great opportunity for researchers in Natural Language Processing, Computer Vision, Machine Learning, Multimodal AI, Symbolic AI, and Data Mining towards the joint understanding of multimedia data on novel abstract concepts.

Comprehensive Semantic Granularity Modeling: Existing models lack the ability to capture various levels of semantic granularity during vision-language pretraining. They rely on human annotation or prior knowledge to control the semantic granularity can be aligned, which is not a satisfactory solution. The ultimate goal is to automatically align semantics of different granularities.

Human-in-the-loop Novel Concept Learning: The goal is to align the human's mental model of the presented scene with the system's model of the scene. As such, the system requests support from the human analysts for low confidence scenes. The interface goals center on different forms of correction with respect to explanations, added examples, quality, and improvements to our representation and curriculum.

Explainable Multimodal Semantics Learning: One exciting research direction is to develop an explainable probabilistic logical rule learning framework. For example, a potential way is to design a vector of probabilistic logical functions on each feature. We then perform probabilistic logical rule learning to learn more expressive rules by composing them. In this way, the model will be able to discover more complex interpretable logical rules from data. Specifically, a family of tree-like logical functions can also be defined. Each logical function will be defined in a recursive manner with logical operations.

6.3.4 Socially-Minded Healthier Information Consumption

Media sources, including social media and contemporary journalism, generate a significant amount of data regarding the exchange of opinions. With the widespread use of multi-modal demonstrations such as images and videos, people tend to frame different narratives based on their preconceived interests. Video framing refers to strategies for narrating videos from a variety of perspectives. By revealing the underlying meaning and bias of the text, it can reveal the author's opinions, intentions, and hidden agendas, which can reduce the level of ambiguity in the text.

Structured knowledge is useful to validate factual knowledge for misinformation detection and to analyze linguistic clues of different framing strategies, such as partial highlights of events, wording, and order of narration. It is an interesting direction to identify writers' opinions, intentions and hidden agendas, thereby reducing ambiguity in text by revealing any underlying meaning and bias. There are immense opportunities for researchers to collaborate with Computing Social Science in order to advance this field.

This direction is also closely related to debiasing foundation models, which have drastically advanced the forefront of various downstream tasks, but the social and cultural commonsense of these models is still an open question. The training data for large language models ranges from obvious discrimination to casual social stereotypes and subtle biases, meaning that these models easily pick up on these negative associations while learning how to effectively use language. This is a rapidly evolving field to recognize the associations that are problematic [377, 378, 379]. However, current approaches mostly focus on post-hoc debiasing, which requires additional mechanisms to effectively handle specific instances of bias, particularly in rare or emerging examples. Another direction worth exploring is enhancing data quality for pretraining by intentionally controlling and mitigating undesired associations, thereby enabling a more informed and nuanced approach.

REFERENCES

[1] M. Stephens, *The Rise of the Image, The Fall of the Word*. New York: Oxford University Press, 1998.

[2] T. Gupta, A. Kamath, A. Kembhavi, and D. Hoiem, "Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01591 pp. 16 378–16 388.

[3] A. Kamath, C. Clark, T. Gupta, E. Kolve, D. Hoiem, and A. Kembhavi, "Webly supervised concept expansion for general purpose vision models," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*. Springer, 2022, pp. 662–681.

[4] M. Li, A. Zareian, Q. Zeng, S. Whitehead, D. Lu, H. Ji, and S.-F. Chang, "Cross-media structured common space for multimedia event extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.acl-main.230 pp. 2557–2568.

[5] Manling Li, R. Xu, S. Wang, X. Lin, C. Zhu, X. Huang, H. Ji, and S.-F. Chang, "Clip-event: Connecting vision and text with event structures," *CVPR*, 2022.

[6] Z. Wang, Manling Li, R. Xu, L. Zhou, J. Lei, X. Lin, S. Wang, Z. Yang, C. Zhu, D. Hoiem, S.-F. Chang, M. Bansal, and H. Ji, "Language models with image descriptors are strong few-shot video-language learners," *NeurIPS*, 2022 (equal contribution).

[7] M. Li, A. Zareian, Y. Lin, X. Pan, S. Whitehead, B. Chen, B. Wu, H. Ji, S.-F. Chang, C. Voss, D. Napierski, and M. Freedman, "GAIA: A fine-grained multimedia knowledge extraction system," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.acl-demos.11 pp. 77–86.

[8] M. Li, Y. Lin, J. Hoover, S. Whitehead, C. Voss, M. Dehghani, and H. Ji, "Multilingual entity, relation, event and human value extraction," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/N19-4019 pp. 110–115.

[9] M. Li, Y. Lin, A. Subburathinam, S. Whitehead et al., "Gaia at sm-kbp 2019-a multi-media multi-lingual knowledge extraction and hypothesis generation system," in *TAC KBP*, 2019.

[10] M. Li, Y. Lin, T. M. Lai, X. Pan et al., "Gaia at sm-kbp 2020 - a dockerized multi-media multi-lingual knowledge extraction, clustering, temporal tracking and hypothesis generation system," in *TAC KBP*, 2020.

[11] Q. Wang, M. Li, X. Wang, N. Parulian, G. Han, J. Ma, J. Tu, Y. Lin, R. H. Zhang, W. Liu, A. Chauhan, Y. Guan, B. Li, R. Li, X. Song, Y. Fung, H. Ji, J. Han, S.-F. Chang, J. Pustejovsky, J. Rah, D. Liem, A. ELsayed, M. Palmer, C. Voss, C. Schneider, and B. Onyshkevych, "COVID-19 literature knowledge graph construction and drug repurposing report generation," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*. Online: Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.naacl-demos.8 pp. 66–77.

[12] OpenAI, "Gpt-4 technical report," *ArXiv preprint*, vol. abs/2303.08774, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[13] M. Li, Q. Zeng, Y. Lin, K. Cho, H. Ji, J. May, N. Chambers, and C. Voss, "Connecting the dots: Event graph schema induction with path language modeling," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.emnlp-main.50 pp. 684–695.

[14] M. Li, S. Li, Z. Wang, L. Huang, K. Cho, H. Ji, J. Han et al., "The future is not one-dimensional: Graph modeling based complex event schema induction for event prediction," *EMNLP*, 2021.

[15] X. Jin, M. Li, and H. Ji, "Event schema induction with double graph autoencoders," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, 2022. [Online]. Available: https://aclanthology.org/2022.naacl-main.147 pp. 2013–2025.

[16] R. G. Reddy, Y. R. Fung, Q. Zeng, M. Li, Z. Wang, P. Sullivan et al., "Smartbook: Ai-assisted situation report generation," *ArXiv preprint*, vol. abs/2303.14337, 2023. [Online]. Available: https://arxiv.org/abs/2303.14337

[17] Y. Guang, Manling Li, J. Zhang, X. Lin, S.-F. Chang, and H. Ji, "Video event extraction via tracking visual states of arguments," *AAAI*, 2023.

[18] S. Li, Z. Wang, M. Li, K. Cho, H. Ji, and j. Han, "Hierarchical event graph schema induction with episode discovery," *Under Review*, 2021.

[19] Q. Zeng, M. Li, T. Lai, H. Ji, M. Bansal, and H. Tong, "GENE: Global event network embedding," in *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*. Mexico City, Mexico: Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.textgraphs-1.5 pp. 42–53.

[20] M. Li, T. Ma, M. Yu, L. Wu, T. Gao, H. Ji, and K. McKeown, "Timeline summarization based on event graph compression via time-aware optimal transport," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.emnlp-main.519 pp. 6443–6456.

[21] M. Li, T. Ma, M. Yu, and A. Fokoue, "Unsupervised knowledge graph compression based on optimal transport," *U.S. Patent submission*, 2020.

[22] M. Li, L. Zhang, H. Ji, and R. J. Radke, "Keep meeting summaries on topic: Abstractive multi-modal meeting summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/P19-1210 pp. 2190–2196.

[23] R. G. Reddy, X. Rui, M. Li, X. Lin, H. Wen, J. Cho, L. Huang, M. Bansal, A. Sil, S.-F. Chang, A. Schwing, and H. Ji, "Multi-media multi-hop news question answering via cross-media grounding," *Under Review*, 2021.

[24] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html pp. 13–23.

[25] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/D19-1514 pp. 5100–5111.

[26] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[27] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei et al., "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *European Conference on Computer Vision*. Springer, 2020, pp. 121–137.

[28] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/7005 pp. 13 041–13 049.

[29] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Revisiting visual representations in vision-language models," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_VinVL_Revisiting_Visual_Representations_in_Vision-Language_Models_CVPR_2021_paper.html pp. 5579–5588.

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021. [Online]. Available: http://proceedings.mlr.press/v139/radford21a.html pp. 8748–8763.

[31] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021. [Online]. Available: http://proceedings.mlr.press/v139/jia21b.html pp. 4904–4916.

[32] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021. [Online]. Available: http://proceedings.mlr.press/v139/kim21k.html pp. 5583–5594.

[33] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. [Online]. Available: https://openreview.net/forum?id=GUrhfTuf_3

[34] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang, "Ernie-vil: Knowledge enhanced vision-language representations through scene graphs," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16431 pp. 3208–3216.

[35] J. Li, R. R. Selvaraju, A. Gotmare, S. R. Joty, C. Xiong, and S. C. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/505259756244493872b7709a8a01b536-Abstract.html pp. 9694–9705.

[36] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li et al., "Florence: A new foundation model for computer vision," *ArXiv preprint*, vol. abs/2111.11432, 2021. [Online]. Available: https://arxiv.org/abs/2111.11432

[37] Z. Huang, Z. Zeng, Y. Huang, B. Liu, D. Fu, and J. Fu, "Seeing out of the box: End-to-end pre-training for vision-language representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Huang_Seeing_Out_of_the_Box_End-to-End_Pre-Training_for_Vision-Language_Representation_CVPR_2021_paper.html pp. 12 976–12 985.

[38] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022. [Online]. Available: https://proceedings.mlr.press/v162/li22n.html pp. 12 888–12 900.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html pp. 5998–6008.

[40] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *ArXiv preprint*, vol. abs/1908.03557, 2019. [Online]. Available: https://arxiv.org/abs/1908.03557

[41] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/6795 pp. 11 336–11 344.

[42] L. H. Li, H. You, Z. Wang, A. Zareian, S.-F. Chang, and K.-W. Chang, "Unsupervised vision-and-language pre-training without parallel images and captions," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.naacl-main.420 pp. 5339–5350.

[43] Z. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng, Z. Liu, and M. Zeng, "An empirical study of training end-to-end vision-and-language transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01763 pp. 18 145–18 155.

[44] A. Miech, J. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00990 pp. 9876–9886.

[45] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som et al., "Image as a foreign language: Beit pretraining for all vision and vision-language tasks," *ArXiv preprint*, vol. abs/2208.10442, 2022. [Online]. Available: https://arxiv.org/abs/2208.10442

[46] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-bert: Aligning image pixels with text by deep multi-modal transformers," *ArXiv preprint*, vol. abs/2004.00849, 2020. [Online]. Available: https://arxiv.org/abs/2004.00849

[47] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. B. Girshick, "Masked autoencoders are scalable vision learners," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01553 pp. 15 979–15 988.

[48] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds et al., "Flamingo: a visual language model for few-shot learning," *ArXiv preprint*, vol. abs/2204.14198, 2022. [Online]. Available: https://arxiv.org/abs/2204.14198

[49] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," *ArXiv preprint*, vol. abs/2202.03052, 2022. [Online]. Available: https://arxiv.org/abs/2202.03052

[50] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *ArXiv preprint*, vol. abs/2301.12597, 2023. [Online]. Available: https://arxiv.org/abs/2301.12597

[51] Y. Yao, A. Zhang, Z. Zhang, Z. Liu, T.-S. Chua, and M. Sun, "Cpt: Colorful prompt tuning for pre-trained vision-language models," *ArXiv preprint*, vol. abs/2109.11797, 2021. [Online]. Available: https://arxiv.org/abs/2109.11797

[52] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *ArXiv preprint*, vol. abs/2109.01134, 2021. [Online]. Available: https://arxiv.org/abs/2109.01134

[53] D. Li, J. Li, H. Li, J. C. Niebles, and S. C. H. Hoi, "Align and prompt: Video-and-language pre-training with entity prompts," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.00490 pp. 4943–4953.

[54] J. Cho, J. Lei, H. Tan, and M. Bansal, "Unifying vision-and-language tasks via text generation," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021. [Online]. Available: http://proceedings.mlr.press/v139/cho21a.html pp. 1931–1942.

[55] M. Tsimpoukelli, J. Menick, S. Cabi, S. M. A. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/01b7575c38dac42f3cfb7d500438b875-Abstract.html pp. 200–212.

[56] Y. Su, T. Lan, Y. Liu, F. Liu, D. Yogatama, Y. Wang, L. Kong, and N. Collier, "Language models can see: Plugging visual controls in text generation," *ArXiv preprint*, vol. abs/2205.02655, 2022. [Online]. Available: https://arxiv.org/abs/2205.02655

[57] X. Zhu, J. Zhu, H. Li, X. Wu, H. Li, X. Wang, and J. Dai, "Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01630 pp. 16 783–16 794.

[58] Z. Wang, M. Li, R. Xu, L. Zhou, J. Lei, X. Lin, S. Wang, Z. Yang, C. Zhu, D. Hoiem et al., "Language models with image descriptors are strong few-shot video-language learners," *ArXiv preprint*, vol. abs/2205.10747, 2022. [Online]. Available: https://arxiv.org/abs/2205.10747

[59] A. Zeng, A. Wong, S. Welker, K. Choromanski, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke et al., "Socratic models: Composing zero-shot multimodal reasoning with language," *ArXiv preprint*, vol. abs/2204.00598, 2022. [Online]. Available: https://arxiv.org/abs/2204.00598

[60] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang, "An empirical study of GPT-3 for few-shot knowledge-based VQA," in *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/20215 pp. 3081–3089.

[61] X. Lin, G. Bertasius, J. Wang, S. Chang, D. Parikh, and L. Torresani, "Vx2text: End-to-end learning of video-based text generation from multimodal inputs," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Lin_Vx2Text_End-to-End_Learning_of_Video-Based_Text_Generation_From_Multimodal_Inputs_CVPR_2021_paper.html pp. 7005–7015.

[62] L. A. Hendricks and A. Nematzadeh, "Probing image-language transformers for verb understanding," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.findings-acl.318 pp. 3635–3644.

[63] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou, "When and why vision-language models behave like bags-of-words, and what to do about it?" in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=KRLUvxh8uaX

[64] J. Lei, T. L. Berg, and M. Bansal, "Revealing single frame bias for video-and-language learning," *ArXiv preprint*, vol. abs/2206.03428, 2022. [Online]. Available: https://arxiv.org/abs/2206.03428

[65] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016. [Online]. Available: https://doi.org/10.1109/CVPR.2016.571 pp. 5288–5296.

[66] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298698 pp. 961–970.

[67] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang, "Ernie-vil: Knowledge enhanced vision-language representations through scene graph," *ArXiv preprint*, vol. abs/2006.16934, 2020. [Online]. Available: https://arxiv.org/abs/2006.16934

[68] L. Chen, Z. Gan, Y. Cheng, L. Li, L. Carin, and J. Liu, "Graph optimal transport for cross-domain alignment," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020. [Online]. Available: http://proceedings.mlr.press/v119/chen20e.html pp. 1542–1553.

[69] A. Zareian, S. Karaman, and S. Chang, "Weakly supervised visual semantic parsing," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00379 pp. 3733–3742.

[70] L. Momeni, M. Caron, A. Nagrani, A. Zisserman, and C. Schmid, "Verbs in action: Improving verb understanding in video-language models," *ArXiv preprint*, vol. abs/2304.06708, 2023. [Online]. Available: https://arxiv.org/abs/2304.06708

[71] Z. Wang, A. Blume, S. Li, G. Liu, J. Cho, Z. Tang, M. Bansal, and H. Ji, "Paxion: Patching action knowledge in video-language foundation models," *ArXiv preprint*, vol. abs/2305.10683, 2023. [Online]. Available: https://arxiv.org/abs/2305.10683

[72] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *ArXiv preprint*, vol. abs/2304.10592, 2023. [Online]. Available: https://arxiv.org/abs/2304.10592

[73] X. Lin, S. Tiwari, S. Huang, M. Li, M. Z. Shou, H. Ji, and S.-F. Chang, "Towards fast adaptation of pretrained contrastive models for multi-channel video-language retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 846–14 855.

[74] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017. [Online]. Available: https://doi.org/10.1109/ICCV.2017.524 pp. 4904–4912.

[75] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 652–663, 2016.

[76] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016. [Online]. Available: https://doi.org/10.1109/CVPR.2016.503 pp. 4651–4659.

[77] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017. [Online]. Available: https://doi.org/10.1109/CVPR.2017.131 pp. 1179–1195.

[78] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: visual question answering," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015. [Online]. Available: https://doi.org/10.1109/ICCV.2015.279 pp. 2425–2433.

[79] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.html pp. 6077–6086.

[80] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017. [Online]. Available: https://doi.org/10.1109/CVPR.2017.670 pp. 6325–6334.

[81] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html pp. 6700–6709.

[82] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Zellers_Neural_Motifs_Scene_CVPR_2018_paper.html pp. 5831–5840.

[83] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Zellers_From_Recognition_to_Cognition_Visual_Commonsense_Reasoning_CVPR_2019_paper.html pp. 6720–6731.

[84] M. Yatskar, L. S. Zettlemoyer, and A. Farhadi, "Situation recognition: Visual semantic role labeling for image understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016. [Online]. Available: https://doi.org/10.1109/CVPR.2016.597 pp. 5534–5542.

[85] R. Li, M. Tapaswi, R. Liao, J. Jia, R. Urtasun, and S. Fidler, "Situation recognition with graph neural networks," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/ICCV.2017.448 pp. 4183–4192.

[86] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fründ, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017. [Online]. Available: https://doi.org/10.1109/ICCV.2017.622 pp. 5843–5851.

[87] G. Gkioxari, R. B. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Gkioxari_Detecting_and_Recognizing_CVPR_2018_paper.html pp. 8359–8367.

[88] J. Lei, L. Yu, M. Bansal, and T. Berg, "TVQA: Localized, compositional video question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018. [Online]. Available: https://aclanthology.org/D18-1167 pp. 1369–1379.

[89] J. Lei, L. Yu, T. Berg, and M. Bansal, "TVQA+: Spatio-temporal grounding for video question answering," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.acl-main.730 pp. 8211–8225.

[90] M. Rohrbach, A. Rohrbach, M. Regneri, S. Amin, M. Andriluka, M. Pinkal, and B. Schiele, "Recognizing fine-grained and composite activities using hand-centric features and script data," *International Journal of Computer Vision*, vol. 119, no. 3, pp. 346–373, 2016.

[91] J. Zhou, K. Lin, H. Li, and W. Zheng, "Graph-based high-order relation modeling for long-term action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Zhou_Graph-Based_High-Order_Relation_Modeling_for_Long-Term_Action_Recognition_CVPR_2021_paper.html pp. 8984–8993.

[92] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, S. A. McIlraith and K. Q. Weinberger, Eds. AAAI Press, 2018. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17344 pp. 7590–7598.

[93] A. Richard, H. Kuehne, and J. Gall, "Action sets: Weakly supervised action segmentation without ordering constraints," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Richard_Action_Sets_Weakly_CVPR_2018_paper.html pp. 5987–5996.

[94] R. Ghoddoosian, S. Sayed, and V. Athitsos, "Hierarchical modeling for task recognition and action segmentation in weakly-labeled instructional videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1922–1932.

[95] D. Zhukov, J. Alayrac, R. G. Cinbis, D. F. Fouhey, I. Laptev, and J. Sivic, "Cross-task weakly supervised learning from instructional videos," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Zhukov_Cross-Task_Weakly_Supervised_Learning_From_Instructional_Videos_CVPR_2019_paper.html pp. 3537–3545.

[96] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou, "COIN: A large-scale dataset for comprehensive instructional video analysis," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Tang_COIN_A_Large-Scale_Dataset_for_Comprehensive_Instructional_Video_Analysis_CVPR_2019_paper.html pp. 1207–1216.

[97] J. Lei, L. Yu, T. Berg, and M. Bansal, "What is more likely to happen next? video-and-language future event prediction," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.emnlp-main.706 pp. 8769–8784.

[98] S. Yang, D. Zha, and C. Xue, "Msk-net: Multi-source knowledge base enhanced networks for script event prediction," in *International Conference on Neural Information Processing*. Springer, 2022, pp. 64–76.

[99] C.-Y. Chang, D.-A. Huang, D. Xu, E. Adeli, L. Fei-Fei, and J. C. Niebles, "Procedure planning in instructional videos," in *European Conference on Computer Vision*. Springer, 2020, pp. 334–350.

[100] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*. IEEE, 1998, pp. 555–562.

[101] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015. [Online]. Available: https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html pp. 91–99.

[102] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016. [Online]. Available: https://doi.org/10.1109/CVPR.2016.91 pp. 779–788.

[103] E. F. Tjong Kim Sang, "Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition," in *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002. [Online]. Available: https://aclanthology.org/W02-2024

[104] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017. [Online]. Available: https://doi.org/10.1109/CVPR.2017.330 pp. 3097–3106.

[105] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *European conference on computer vision*. Springer, 2016, pp. 852–869.

[106] A. Culotta and J. Sorensen, "Dependency tree kernels for relation extraction," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, Barcelona, Spain, 2004. [Online]. Available: https://aclanthology.org/P04-1054 pp. 423–429.

[107] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "VL-BERT: pre-training of generic visual-linguistic representations," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=SygXPaEYvH

[108] Y. Liu, C. Wu, S.-Y. Tseng, V. Lal, X. He, and N. Duan, "KD-VLP: Improving end-to-end vision-and-language pretraining with object knowledge distillation," in *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, 2022. [Online]. Available: https://aclanthology.org/2022.findings-naacl.119 pp. 1589–1600.

[109] R. Cadène, H. Ben-younes, M. Cord, and N. Thome, "MUREL: multimodal relational reasoning for visual question answering," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Cadene_MUREL_Multimodal_Relational_Reasoning_for_Visual_Question_Answering_CVPR_2019_paper.html pp. 1989–1998.

[110] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019. [Online]. Available: https://doi.org/10.1109/ICCV.2019.01041 pp. 10 312–10 321.

[111] M. Li, R. Xu, S. Wang, L. Zhou, X. Lin, C. Zhu, M. Zeng, H. Ji, and S. Chang, "Clip-event: Connecting text and images with event structures," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01593 pp. 16 399–16 408.

[112] A. Sadhu, T. Gupta, M. Yatskar, R. Nevatia, and A. Kembhavi, "Visual semantic role labeling for video understanding," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Sadhu_Visual_Semantic_Role_Labeling_for_Video_Understanding_CVPR_2021_paper.html pp. 5589–5600.

[113] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298698 pp. 961–970.

[114] B. Yao and F. Li, "Modeling mutual context of object and human pose in human-object interaction activities," in *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*. IEEE Computer Society, 2010. [Online]. Available: https://doi.org/10.1109/CVPR.2010.5540235 pp. 17–24.

[115] L. Zhu and Y. Yang, "Actbert: Learning global-local video-text representations," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00877 pp. 8743–8752.

[116] R. C. Schank, *Dynamic memory: A theory of reminding and learning in computers and people*. cambridge university press, 1983.

[117] wikiHow, "wikiHow," https://www.wikiHow.com/.

[118] J. Bi, J. Luo, and C. Xu, "Procedure planning in instructional videos via contextual modeling and model-based policy learning," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.01532 pp. 15 591–15 600.

[119] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price et al., "Scaling egocentric vision: The epic-kitchens dataset," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 720–736.

[120] A. Furnari and G. M. Farinella, "What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00635 pp. 6251–6260.

[121] W. Lotter, G. Kreiman, and D. D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=B1ewdt9xe

[122] T. Han, W. Xie, and A. Zisserman, "Video representation learning by dense predictive coding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[123] F. Sener and A. Yao, "Zero-shot anticipation for instructional activities," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00095 pp. 862–871.

[124] X. Lin, F. Petroni, G. Bertasius, M. Rohrbach, S. Chang, and L. Torresani, "Learning to recognize procedural activities with distant supervision," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01348 pp. 13 843–13 853.

[125] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "MERLOT: multimodal neural script knowledge models," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/c6d4eb15f1e84a36eff58eca3627c82e-Abstract.html pp. 23 634–23 651.

[126] R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, and Y. Choi, "Merlot reserve: Neural script knowledge through vision and language and sound," *ArXiv preprint*, vol. abs/2201.02639, 2022. [Online]. Available: https://arxiv.org/abs/2201.02639

[127] F. F. Xu, L. Ji, B. Shi, J. Du, G. Neubig, Y. Bisk, and N. Duan, "A benchmark for structured procedural knowledge extraction from cooking videos," in *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.nlpbt-1.4 pp. 30–40.

[128] Y. Yang, J. Kim, A. Panagopoulou, M. Yatskar, and C. Callison-Burch, "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval," *ArXiv preprint*, vol. abs/2111.09276, 2021. [Online]. Available: https://arxiv.org/abs/2111.09276

[129] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/N19-1423 pp. 4171–4186.

[130] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[131] C. Li, H. Liu, L. H. Li, P. Zhang, J. Aneja, J. Yang, P. Jin, Y. J. Lee, H. Hu, Z. Liu et al., "Elevater: A benchmark and toolkit for evaluating language-augmented visual models," *ArXiv preprint*, vol. abs/2204.08790, 2022. [Online]. Available: https://arxiv.org/abs/2204.08790

[132] X. Lin, S. Tiwari, S. Huang, M. Li, M. Z. Shou, H. Ji, and S.-F. Chang, "Towards fast adaptation of pretrained contrastive models for multi-channel video-language retrieval," *ArXiv preprint*, vol. abs/2206.02082, 2022. [Online]. Available: https://arxiv.org/abs/2206.02082

[133] C. Zhang, B. Van Durme, Z. Li, and E. Stengel-Eskin, "Visual commonsense in pretrained unimodal and multimodal models," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, 2022. [Online]. Available: https://aclanthology.org/2022.naacl-main.390 pp. 5321–5335.

[134] H. Xue, Y. Huang, B. Liu, H. Peng, J. Fu, H. Li, and J. Luo, "Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/23fa71cc32babb7b91130824466d25a5-Abstract.html pp. 4514–4528.

[135] K. Chen, Q. Huang, D. McDuff, Y. Bisk, and J. Gao, "Krit: Knowledge-reasoning intelligence in vision-language transformer," 2022.

[136] H. Tan and M. Bansal, "Vokenization: Improving language understanding with contextualized, visual-grounded supervision," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.emnlp-main.162 pp. 2066–2080.

[137] Z. Tang, J. Cho, H. Tan, and M. Bansal, "Vidlankd: Improving language understanding via video-distilled knowledge transfer," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/ccdf3864e2fa9089f9eca4fc7a48ea0a-Abstract.html pp. 24 468–24 481.

[138] A. P. D. Mourelatos, "Events, processes, and states," *Linguistics and Philosophy*, vol. 2, pp. 415–434, 1978.

[139] E. Bach, "The algebra of events," *Linguistics and philosophy*, vol. 9, no. 1, pp. 5–16, 1986.

[140] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet project," in *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998. [Online]. Available: https://aclanthology.org/C98-1013

[141] H. Zhang, X. Liu, H. Pan, Y. Song, and C. W. Leung, "ASER: A large-scale eventuality knowledge graph," in *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Y. Huang, I. King, T. Liu, and M. van Steen, Eds. ACM / IW3C2, 2020. [Online]. Available: https://doi.org/10.1145/3366423.3380107 pp. 201–211.

[142] N. Chambers and D. Jurafsky, "Unsupervised learning of narrative event chains," in *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, 2008. [Online]. Available: https://aclanthology.org/P08-1090 pp. 789–797.

[143] H. Peng, S. Chaturvedi, and D. Roth, "A joint model for semantic sequences: Frames, entities, sentiments," in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Vancouver, Canada: Association for Computational Linguistics, 2017. [Online]. Available: https://aclanthology.org/K17-1019 pp. 173–183.

[144] L. He, M. Lewis, and L. Zettlemoyer, "Question-answer driven semantic role labeling: Using natural language to annotate natural language," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015. [Online]. Available: https://aclanthology.org/D15-1076 pp. 643–653.

[145] J. Michael, G. Stanovsky, L. He, I. Dagan, and L. Zettlemoyer, "Crowdsourcing question-answer meaning representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018. [Online]. Available: https://aclanthology.org/N18-2089 pp. 560–568.

[146] T. Goyal and G. Durrett, "Embedding time expressions for deep temporal ordering models," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/P19-1433 pp. 4400–4406.

[147] H. Ji and R. Grishman, "Refining event extraction through cross-document inference," in *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, 2008. [Online]. Available: https://aclanthology.org/P08-1030 pp. 254–262.

[148] S. Liao and R. Grishman, "Acquiring topic features to improve event extraction: in pre-selected and balanced collections," in *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Hissar, Bulgaria: Association for Computational Linguistics, 2011. [Online]. Available: https://aclanthology.org/R11-1002 pp. 9–16.

[149] R. Huang and E. Riloff, "Bootstrapped training of event extraction classifiers," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, 2012. [Online]. Available: https://aclanthology.org/E12-1029 pp. 286–295.

[150] Y. Chen, L. Xu, K. Liu, D. Zeng, and J. Zhao, "Event extraction via dynamic multi-pooling convolutional neural networks," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, 2015. [Online]. Available: https://aclanthology.org/P15-1017 pp. 167–176.

[151] T. H. Nguyen, K. Cho, and R. Grishman, "Joint event extraction via recurrent neural networks," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, 2016. [Online]. Available: https://aclanthology.org/N16-1034 pp. 300–309.

[152] Y. Hong, W. Zhou, J. Zhang, G. Zhou, and Q. Zhu, "Self-regulation: Employing a generative adversarial network to improve event detection," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018. [Online]. Available: https://aclanthology.org/P18-1048 pp. 515–526.

[153] X. Liu, Z. Luo, and H. Huang, "Jointly multiple events extraction via attention-based graph information aggregation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018. [Online]. Available: https://aclanthology.org/D18-1156 pp. 1247–1256.

[154] Y. Chen, H. Yang, K. Liu, J. Zhao, and Y. Jia, "Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018. [Online]. Available: https://aclanthology.org/D18-1158 pp. 1267–1276.

[155] T. Zhang, H. Ji, and A. Sil, "Joint entity and event extraction with generative adversarial imitation learning," *Data Intelligence Vol 1 (2): 99-120*, 2019.

[156] X. Liu, Z. Luo, and H. Huang, "Jointly multiple events extraction via attention-based graph information aggregation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018. [Online]. Available: https://aclanthology.org/D18-1156 pp. 1247–1256.

[157] R. Wang, D. Zhou, and Y. He, "Open event extraction from online text using a generative adversarial network," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/D19-1027 pp. 282–291.

[158] S. Yang, D. Feng, L. Qiao, Z. Kan, and D. Li, "Exploring pre-trained language models for event extraction and generation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/P19-1522 pp. 5284–5294.

[159] Q. Li, H. Ji, and L. Huang, "Joint event extraction via structured prediction with global features," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, 2013. [Online]. Available: https://aclanthology.org/P13-1008 pp. 73–82.

[160] B. Yang and T. M. Mitchell, "Joint extraction of events and entities within a document context," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, 2016. [Online]. Available: https://aclanthology.org/N16-1033 pp. 289–299.

[161] D. Wadden, U. Wennberg, Y. Luan, and H. Hajishirzi, "Entity, relation, and event extraction with contextualized span representations," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/D19-1585 pp. 5784–5789.

[162] Y. Lin, H. Ji, F. Huang, and L. Wu, "A joint neural model for information extraction with global features," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.acl-main.713 pp. 7999–8009.

[163] Q. Lyu, H. Zhang, E. Sulem, and D. Roth, "Zero-shot event extraction via transfer learning: Challenges and insights," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.acl-short.42 pp. 322–332.

[164] O. Levy, M. Seo, E. Choi, and L. Zettlemoyer, "Zero-shot relation extraction via reading comprehension," in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Vancouver, Canada: Association for Computational Linguistics, 2017. [Online]. Available: https://aclanthology.org/K17-1034 pp. 333–342.

[165] X. Li, F. Yin, Z. Sun, X. Li, A. Yuan, D. Chai, M. Zhou, and J. Li, "Entity-relation extraction as multi-turn question answering," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/P19-1129 pp. 1340–1350.

[166] X. Du and C. Cardie, "Event extraction by answering (almost) natural questions," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.emnlp-main.49 pp. 671–683.

[167] B. Li, W. Yin, and M. Chen, "Ultra-fine entity typing with indirect supervision from natural language inference," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 607–622, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.35

[168] W. Yin, N. F. Rajani, D. Radev, R. Socher, and C. Xiong, "Universal natural language processing with limited annotations: Try few-shot textual entailment as a start," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.emnlp-main.660 pp. 8229–8239.

[169] Y. Lu, H. Lin, J. Xu, X. Han, J. Tang, A. Li, L. Sun, M. Liao, and S. Chen, "Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.acl-long.217 pp. 2795–2806.

[170] S. Li, H. Ji, and J. Han, "Document-level event argument extraction by conditional generation," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.naacl-main.69 pp. 894–908.

[171] T. Zhang, S. Whitehead, H. Zhang, H. Li, J. G. Ellis, L. Huang, W. Liu, H. Ji, and S. Chang, "Improving event extraction via multimodal integration," in *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, 2017. [Online]. Available: https://doi.org/10.1145/3123266.3123294 pp. 270–278.

[172] X. Chang, Z. Ma, Y. Yang, Z. Zeng, and A. G. Hauptmann, "Bi-level semantic representation analysis for multimedia event detection," *IEEE transactions on cybernetics*, vol. 47, no. 5, pp. 1180–1197, 2016.

[173] Y. Zhang, C. Xu, Y. Rui, J. Wang, and H. Lu, "Semantic event extraction from basketball games using multi-modal analysis," in *2007 IEEE International Conference on Multimedia and Expo*. IEEE, 2007, pp. 2190–2193.

[174] Z. Ma, X. Chang, Z. Xu, N. Sebe, and A. G. Hauptmann, "Joint attributes and event analysis for multimedia event detection," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 7, pp. 2921–2930, 2017.

[175] A. A. Perera, S. Oh, P. Megha, T. Ma, A. Hoogs, A. Vahdat, K. Cannons, G. Mori, S. Mccloskey, B. Miller et al., "Trecvid 2012 genie: Multimedia event detection and recounting," in *In TRECVID Workshop*. Citeseer, 2012.

[176] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Gu_AVA_A_Video_CVPR_2018_paper.html pp. 6047–6056.

[177] D. Li, Z. Qiu, Q. Dai, T. Yao, and T. Mei, "Recurrent tubelet proposal and recognition networks for action detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 303–318.

[178] K. Duarte, Y. S. Rawat, and M. Shah, "Videocapsulenet: A simplified network for action detection," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/73f104c9fba50050eea11d9d075247cc-Abstract.html pp. 7621–7630.

[179] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European Conference on Computer Vision*. Springer, 2016, pp. 510–526.

[180] K. Kato, Y. Li, and A. Gupta, "Compositional learning for human object interaction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 234–251.

[181] C. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krähenbühl, and R. B. Girshick, "Long-term feature banks for detailed video understanding," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Wu_Long-Term_Feature_Banks_for_Detailed_Video_Understanding_CVPR_2019_paper.html pp. 284–293.

[182] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *2012 IEEE international conference on robotics and automation*. IEEE, 2012, pp. 842–849.

[183] Y. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H. Fang, Y. Wang, and C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Li_Transferable_Interactiveness_Knowledge_for_Human-Object_Interaction_Detection_CVPR_2019_paper.html pp. 3585–3594.

[184] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00417 pp. 4115–4124.

[185] T. Zhou, W. Wang, S. Qi, H. Ling, and J. Shen, "Cascaded human-object interaction recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00432 pp. 4262–4271.

[186] A. Mallya and S. Lazebnik, "Recurrent models for situation recognition," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017. [Online]. Available: https://doi.org/10.1109/ICCV.2017.57 pp. 455–463.

[187] S. Pratt, M. Yatskar, L. Weihs, A. Farhadi, and A. Kembhavi, "Grounded situation recognition," in *European Conference on Computer Vision*. Springer, 2020, pp. 314–332.

[188] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet project," in *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998. [Online]. Available: https://aclanthology.org/C98-1013

[189] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," in *European conference on computer vision*. Springer, 2010, pp. 494–507.

[190] H. Kuehne, H. Jhuang, E. Garrote, T. A. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, D. N. Metaxas, L. Quan, A. Sanfeliu, and L. V. Gool, Eds. IEEE Computer Society, 2011. [Online]. Available: https://doi.org/10.1109/ICCV.2011.6126543 pp. 2556–2563.

[191] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *ECCV*, 2016.

[192] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017. [Online]. Available: https://doi.org/10.1109/CVPR.2017.502 pp. 4724–4733.

[193] Z. Shou, X. Lin, Y. Kalantidis, L. Sevilla-Lara, M. Rohrbach, S. Chang, and Z. Yan, "Dmc-net: Generating discriminative motion cues for fast compressed video action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Shou_DMC-Net_Generating_Discriminative_Motion_Cues_for_Fast_Compressed_Video_Action_CVPR_2019_paper.html pp. 1268–1277.

[194] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev et al., "The kinetics human action video dataset," *ArXiv preprint*, vol. abs/1705.06950, 2017. [Online]. Available: https://arxiv.org/abs/1705.06950

[195] F. Baradel, N. Neverova, C. Wolf, J. Mille, and G. Mori, "Object level visual reasoning in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 105–121.

[196] J. C. Stroud, D. A. Ross, C. Sun, J. Deng, and R. Sukthankar, "D3d: Distilled 3d networks for video action recognition," *ArXiv preprint*, vol. abs/1812.08249, 2018. [Online]. Available: https://arxiv.org/abs/1812.08249

[197] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "MARS: motion-augmented RGB stream for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Crasto_MARS_Motion-Augmented_RGB_Stream_for_Action_Recognition_CVPR_2019_paper.html pp. 7882–7891.

[198] K. Gao, L. Chen, Y. Huang, and J. Xiao, "Video relation detection via tracklet based visual transformer," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4833–4837.

[199] K. Gao, L. Chen, Y. Niu, J. Shao, and J. Xiao, "Classification-then-grounding: Reformulating video scene graphs as temporal bipartite graphs," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01889 pp. 19 475–19 484.

[200] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00630 pp. 6201–6210.

[201] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021. [Online]. Available: http://proceedings.mlr.press/v139/bertasius21a.html pp. 813–824.

[202] X. Lin, L. Ma, W. Liu, and S.-F. Chang, "Context-gated convolution," in *European Conference on Computer Vision*. Springer, 2020, pp. 701–718.

[203] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017. [Online]. Available: https://doi.org/10.1109/CVPR.2017.502 pp. 4724–4733.

[204] B. Chen, X. Lin, C. Thomas, M. Li, S. Yoshida, L. Chum, H. Ji, and S.-F. Chang, "Joint multimedia event extraction from video and article," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.8 pp. 74–88.

[205] X. Zhu, Z. Li, X. Wang, X. Jiang, P. Sun, X. Wang, Y. Xiao, and N. J. Yuan, "Multimodal knowledge graph construction and application: A survey," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[206] L. Zhang, D. Zhou, Y. He, and Z. Yang, "MERL: multimodal event representation learning in heterogeneous embedding spaces," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17695 pp. 14 420–14 427.

[207] S. Chaturvedi, H. Peng, and D. Roth, "Story comprehension for predicting what happens next," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017. [Online]. Available: https://aclanthology.org/D17-1168 pp. 1603–1614.

[208] H. Zhang, M. Chen, H. Wang, Y. Song, and D. Roth, "Open-domain process structure induction," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2020.

[209] M. Chen, H. Zhang, H. Wang, and D. Roth, "What are you trying to do? semantic typing of event processes," in *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.conll-1.43 pp. 531–542.

[210] C. Zhu, M. Chen, C. Fan, G. Cheng, and Y. Zhang, "Learning from history: Modeling temporal knowledge graphs with sequential copy-generation networks," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16604 pp. 4732–4740.

[211] A. Cybulska and P. Vossen, "Guidelines for ecb+ annotation of events and their coreference," in *Technical Report*. Technical Report NWR-2014-1, VU University Amsterdam, 2014.

[212] R. C. Schank and R. P. Abelson, "Scripts, plans, goals and understanding: An inquiry into human knowledge structures." *Mhwah, NJ (US): Lawrence Erlbaum Associates*, 1977.

[213] N. Chambers and D. Jurafsky, "Unsupervised learning of narrative event chains," in *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, 2008. [Online]. Available: https://aclanthology.org/P08-1090 pp. 789–797.

[214] N. Balasubramanian, S. Soderland, Mausam, and O. Etzioni, "Generating coherent event schemas at scale," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, 2013. [Online]. Available: https://aclanthology.org/D13-1178 pp. 1721–1731.

[215] N. Chambers, "Event schema induction with a probabilistic entity-driven model," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, 2013. [Online]. Available: https://aclanthology.org/D13-1185 pp. 1797–1807.

[216] J. C. K. Cheung, H. Poon, and L. Vanderwende, "Probabilistic frame induction," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, 2013. [Online]. Available: https://aclanthology.org/N13-1104 pp. 837–846.

[217] K. Pichotta and R. J. Mooney, "Learning statistical scripts with LSTM recurrent neural networks," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, D. Schuurmans and M. P. Wellman, Eds. AAAI Press, 2016. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12157 pp. 2800–2806.

[218] A. Modi and I. Titov, "Inducing neural models of script knowledge," in *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Ann Arbor, Michigan: Association for Computational Linguistics, 2014. [Online]. Available: https://aclanthology.org/W14-1606 pp. 49–57.

[219] R. Rudinger, P. Rastogi, F. Ferraro, and B. Van Durme, "Script induction as language modeling," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015. [Online]. Available: https://aclanthology.org/D15-1195 pp. 1681–1686.

[220] M. Li, Q. Zeng, Y. Lin, K. Cho, H. Ji, J. May, N. Chambers, and C. Voss, "Connecting the dots: Event graph schema induction with path language modeling," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.emnlp-main.50 pp. 684–695.

[221] M. Li, S. Li, Z. Wang, L. Huang, K. Cho, H. Ji, and J. Han, "Future is not one-dimensional: Complex event schema induction via graph modeling," in *arxiv2104.06344*, 2021.

[222] L. Zhang, Q. Lyu, and C. Callison-Burch, "Reasoning about goals, steps, and temporal ordering with WikiHow," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.emnlp-main.374 pp. 4630–4639.

[223] Y. Yang, A. Panagopoulou, Q. Lyu, L. Zhang, M. Yatskar, and C. Callison-Burch, "Visual goal-step inference using wikiHow," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.emnlp-main.165 pp. 2167–2179.

[224] N. Weber, L. Shekhar, H. Kwon, N. Balasubramanian, and N. Chambers, "Generating narrative text in a switching dynamical system," in *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.conll-1.42 pp. 520–530.

[225] X. Ding, K. Liao, T. Liu, Z. Li, and J. Duan, "Event representation learning enhanced with external commonsense knowledge," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/D19-1495 pp. 4894–4903.

[226] N. Weber, R. Rudinger, and B. Van Durme, "Causal inference of script knowledge," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.emnlp-main.612 pp. 7583–7596.

[227] N. Chambers and D. Jurafsky, "Unsupervised learning of narrative schemas and their participants," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore: Association for Computational Linguistics, 2009. [Online]. Available: https://aclanthology.org/P09-1068 pp. 602–610.

[228] N. Chambers and D. Jurafsky, "A database of narrative schemas," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/58_Paper.pdf

[229] B. Jans, S. Bethard, I. Vulić, and M. F. Moens, "Skip n-grams and ranking functions for predicting script events," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, 2012. [Online]. Available: https://aclanthology.org/E12-1034 pp. 336–344.

[230] K. Pichotta and R. Mooney, "Statistical script learning with multi-argument events," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, 2014. [Online]. Available: https://aclanthology.org/E14-1024 pp. 220–229.

[231] M. Granroth-Wilding and S. Clark, "What happens next? event prediction using a compositional neural network model," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, D. Schuurmans and M. P. Wellman, Eds. AAAI Press, 2016. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11995 pp. 2727–2733.

[232] A. Modi, "Event embeddings for semantic script modeling," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany: Association for Computational Linguistics, 2016. [Online]. Available: https://aclanthology.org/K16-1008 pp. 75–83.

[233] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, and J. Allen, "A corpus and cloze evaluation for deeper understanding of commonsense stories," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, 2016. [Online]. Available: https://aclanthology.org/N16-1098 pp. 839–849.

[234] H. Peng, Q. Ning, and D. Roth, "KnowSemLM: A knowledge infused semantic language model," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/K19-1051 pp. 550–562.

[235] Z. Li, X. Ding, and T. Liu, "Constructing narrative event evolutionary graph for script event prediction," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, J. Lang, Ed. ijcai.org, 2018. [Online]. Available: https://doi.org/10.24963/ijcai.2018/584 pp. 4201–4207.

[236] N. Mostafazadeh, A. Grealish, N. Chambers, J. Allen, and L. Vanderwende, "CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures," in *Proceedings of the Fourth Workshop on Events*. San Diego, California: Association for Computational Linguistics, 2016. [Online]. Available: https://aclanthology.org/W16-1007 pp. 51–61.

[237] P. Kalm, M. Regan, and W. Croft, "Event structure representation: Between verbs and argument structure constructions," in *Proceedings of the First International Workshop on Designing Meaning Representations*. Florence, Italy: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/W19-3311 pp. 100–109.

[238] A. Modi, T. Anikina, S. Ostermann, and M. Pinkal, "InScript: Narrative texts annotated with script information," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. [Online]. Available: https://aclanthology.org/L16-1555 pp. 3485–3493.

[239] L. Wanzare, A. Zarcone, S. Thater, and M. Pinkal, "Inducing script structure from crowdsourced event descriptions via semi-supervised clustering," in *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*. Valencia, Spain: Association for Computational Linguistics, 2017. [Online]. Available: https://aclanthology.org/W17-0901 pp. 1–11.

[240] F. Zhai, V. Demberg, P. Shkadzko, W. Shi, and A. Sayeed, "A hybrid model for globally coherent story generation," in *Proceedings of the Second Workshop on Storytelling*. Florence, Italy: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/W19-3404 pp. 34–45.

[241] D. Q. Nguyen, D. Q. Nguyen, C. X. Chu, S. Thater, and M. Pinkal, "Sequence to sequence learning for event prediction," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, 2017. [Online]. Available: https://aclanthology.org/I17-2007 pp. 37–42.

[242] L. Hu, J. Li, L. Nie, X. Li, and C. Shao, "What happens next? future subevent prediction using contextual hierarchical LSTM," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, S. P. Singh and S. Markovitch, Eds. AAAI Press, 2017. [Online]. Available: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14324 pp. 3450–3456.

[243] Z. Li, X. Ding, and T. Liu, "Constructing narrative event evolutionary graph for script event prediction," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, J. Lang, Ed. ijcai.org, 2018. [Online]. Available: https://doi.org/10.24963/ijcai.2018/584 pp. 4201–4207.

[244] H. Kiyomaru, K. Omura, Y. Murawaki, D. Kawahara, and S. Kurohashi, "Diversity-aware event prediction based on a conditional variational autoencoder with reconstruction," in *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*. Hong Kong, China: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/D19-6014 pp. 113–122.

[245] S. Lv, W. Qian, L. Huang, J. Han, and S. Hu, "Sam-net: Integrating event-level and chain-level attentions to predict what happens next," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33016802 pp. 6802–6809.

[246] H. Kwon, M. Koupaee, P. Singh, G. Sawhney, A. Shukla, K. K. Kallur, N. Chambers, and N. Balasubramanian, "Modeling preconditions in text with a crowd-sourced dataset," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.340 pp. 3818–3828.

[247] A. Prakash, M. Blanchette, S. Sinha, and M. Tompa, "Motif discovery in heterogeneous sequence data," in *Biocomputing 2004, Proceedings of the Pacific Symposium, Hawaii, USA, 6-10 January 2004*, R. B. Altman, A. K. Dunker, L. Hunter, T. A. Jung, and T. E. Klein, Eds. World Scientific, 2004. [Online]. Available: http://psb.stanford.edu/psb-online/proceedings/psb04/prakash.pdf pp. 348–359.

[248] A. G. Carranza, R. A. Rossi, A. Rao, and E. Koh, "Higher-order spectral clustering for heterogeneous graphs," *ArXiv preprint*, vol. abs/1810.02959, 2018. [Online]. Available: https://arxiv.org/abs/1810.02959

[249] R. A. Rossi, N. K. Ahmed, A. G. Carranza, D. Arbour, A. Rao, S. Kim, and E. Koh, "Heterogeneous network motifs," *ArXiv preprint*, vol. abs/1901.10026, 2019. [Online]. Available: https://arxiv.org/abs/1901.10026

[250] J. Hu, R. Cheng, K. C. Chang, A. Sankar, Y. Fang, and B. Y. H. Lam, "Discovering maximal motif cliques in large heterogeneous information networks," in *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*. IEEE, 2019. [Online]. Available: https://doi.org/10.1109/ICDE.2019.00072 pp. 746–757.

[251] D. J. Cook and L. B. Holder, "Substructure discovery using minimum description length and background knowledge," *Journal of Artificial Intelligence Research*, vol. 1, pp. 231–255, 1993.

[252] G. Buehrer and K. Chellapilla, "A scalable pattern mining approach to web graph compression with communities," in *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008*, M. Najork, A. Z. Broder, and S. Chakrabarti, Eds. ACM, 2008. [Online]. Available: https://doi.org/10.1145/1341531.1341547 pp. 95–106.

[253] C.-T. Li and S.-D. Lin, "Egocentric information abstraction for heterogeneous social networks," in *2009 International Conference on Advances in Social Network Analysis and Mining*. IEEE, 2009, pp. 255–260.

[254] N. Zhang, Y. Tian, and J. M. Patel, "Discovery-driven graph summarization," in *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*. IEEE, 2010, pp. 880–891.

[255] D. Koutra, U. Kang, J. Vreeken, and C. Faloutsos, "VOG: summarizing and understanding large graphs," in *Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24-26, 2014*, M. J. Zaki, Z. Obradovic, P. Tan, A. Banerjee, C. Kamath, and S. Parthasarathy, Eds. SIAM, 2014. [Online]. Available: https://doi.org/10.1137/1.9781611973440.11 pp. 91–99.

[256] Y. Wu, Z. Zhong, W. Xiong, and N. Jing, "Graph summarization for attributed graphs," in *2014 International Conference on Information Science, Electronics and Electrical Engineering*, vol. 1. IEEE, 2014, pp. 503–507.

[257] Q. Song, Y. Wu, P. Lin, L. X. Dong, and H. Sun, "Mining summaries for knowledge graph search," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 10, pp. 1887–1900, 2018.

[258] F. Bariatti, P. Cellier, and S. Ferré, "Graphmdl: Graph pattern selection based on minimum description length," in *International Symposium on Intelligent Data Analysis*. Springer, 2020, pp. 54–66.

[259] T. Kipf and M. Welling, "Variational graph auto-encoders," *ArXiv preprint*, vol. abs/1611.07308, 2016. [Online]. Available: https://arxiv.org/abs/1611.07308

[260] M. Simonovsky and N. Komodakis, "Graphvae: Towards generation of small graphs using variational autoencoders," in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 412–422.

[261] A. Grover, A. Zweig, and S. Ermon, "Graphite: Iterative generative modeling of graphs," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019. [Online]. Available: http://proceedings.mlr.press/v97/grover19a.html pp. 2434–2444.

[262] J. Liu, A. Kumar, J. Ba, J. Kiros, and K. Swersky, "Graph normalizing flows," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/1e44fdf9c44d7328fecc02d677ed704d-Abstract.html pp. 13 556–13 566.

[263] A. Bojchevski, O. Shchur, D. Zügner, and S. Günnemann, "Netgan: Generating graphs via random walks," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018. [Online]. Available: http://proceedings.mlr.press/v80/bojchevski18a.html pp. 609–618.

[264] D. Zhou, L. Zheng, J. Han, and J. He, "A data-driven graph generative model for temporal interaction networks," in *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, R. Gupta, Y. Liu, J. Tang, and B. A. Prakash, Eds. ACM, 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3394486.3403082 pp. 401–411.

[265] Y. Li, O. Vinyals, C. Dyer, R. Pascanu, and P. Battaglia, "Learning deep generative models of graphs," in *Proceedings of the 35 th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80*, 2018.

[266] J. You, R. Ying, X. Ren, W. L. Hamilton, and J. Leskovec, "Graphrnn: Generating realistic graphs with deep auto-regressive models," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018. [Online]. Available: http://proceedings.mlr.press/v80/you18a.html pp. 5694–5703.

[267] R. Liao, Y. Li, Y. Song, S. Wang, W. L. Hamilton, D. Duvenaud, R. Urtasun, and R. S. Zemel, "Efficient graph generation with graph recurrent attention networks," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/d0921d442ee91b896ad95059d13df618-Abstract.html pp. 4257–4267.

[268] F. Sener and A. Yao, "Zero-shot anticipation for instructional activities," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00095 pp. 862–871.

[269] H. Zhou, R. Mart'in-Mart'in, M. Kapadia, S. Savarese, and J. C. Niebles, "Procedure-aware pretraining for instructional video understanding," *ArXiv preprint*, vol. abs/2303.18230, 2023. [Online]. Available: https://arxiv.org/abs/2303.18230

[270] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "MERLOT: multimodal neural script knowledge models," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/c6d4eb15f1e84a36eff58eca3627c82e-Abstract.html pp. 23 634–23 651.

[271] A. Salvador, M. Drozdzal, X. Giró-i-Nieto, and A. Romero, "Inverse cooking: Recipe generation from food images," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Salvador_Inverse_Cooking_Recipe_Generation_From_Food_Images_CVPR_2019_paper.html pp. 10 453–10 462.

[272] Y. Yang, J. Kim, A. Panagopoulou, M. Yatskar, and C. Callison-Burch, "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval," *ArXiv preprint*, vol. abs/2111.09276, 2021. [Online]. Available: https://arxiv.org/abs/2111.09276

[273] L. Logeswaran, S. Sohn, Y. Jang, M. Lee, and H. H. Lee, "Unsupervised task graph generation from instructional video transcripts," *ArXiv preprint*, vol. abs/2302.09173, 2023. [Online]. Available: https://arxiv.org/abs/2302.09173

[274] Y. Jang, S. Sohn, L. Logeswaran, T. Luo, M. Lee, and H. H. Lee, "Multimodal subtask graph generation from instructional videos," *ArXiv preprint*, vol. abs/2302.08672, 2023. [Online]. Available: https://arxiv.org/abs/2302.08672

[275] J. Bi, J. Luo, and C. Xu, "Procedure planning in instructional videos via contextual modeling and model-based policy learning," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.01532 pp. 15 591–15 600.

[276] A. Srinivas, A. Jabri, P. Abbeel, S. Levine, and C. Finn, "Universal planning networks: Learning generalizable representations for visuomotor control," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018. [Online]. Available: http://proceedings.mlr.press/v80/srinivas18b.html pp. 4739–4748.

[277] J. Sun, D.-A. Huang, B. Lu, Y.-H. Liu, B. Zhou, and A. Garg, "Plate: Visually-grounded planning with transformers in procedural tasks," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4924–4930, 2022.

[278] H. Zhao, I. Hadji, N. Dvornik, K. G. Derpanis, R. P. Wildes, and A. D. Jepson, "P3iv: Probabilistic procedure planning from instructional videos with weak supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2938–2948.

[279] M.-A. Côté, A. Kádár, X. Yuan, B. Kybartas, T. Barnes, E. Fine, J. Moore, M. Hausknecht, L. E. Asri, M. Adada et al., "Textworld: A learning environment for text-based games," in *Workshop on Computer Games*. Springer, 2018, pp. 41–75.

[280] M. Shridhar, X. Yuan, M. Côté, Y. Bisk, A. Trischler, and M. J. Hausknecht, "Alfworld: Aligning text and embodied environments for interactive learning," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=0IOX0YcCdTn

[281] V. Micheli and F. Fleuret, "Language models are few-shot butlers," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.emnlp-main.734 pp. 9312–9318.

[282] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. C. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. M. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, and M. Yan, "Do as i can, not as i say: Grounding language in robotic affordances," *ArXiv preprint*, vol. abs/2204.01691, 2022. [Online]. Available: https://arxiv.org/abs/2204.01691

[283] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar et al., "Inner monologue: Embodied reasoning through planning with language models," *ArXiv preprint*, vol. abs/2207.05608, 2022. [Online]. Available: https://arxiv.org/abs/2207.05608

[284] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," *ArXiv preprint*, vol. abs/2210.03629, 2022. [Online]. Available: https://arxiv.org/abs/2210.03629

[285] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022. [Online]. Available: https://proceedings.mlr.press/v162/huang22a.html pp. 9118–9147.

[286] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study final report," 1998.

[287] P. Laban and M. Hearst, "newsLens: building and visualizing long-ranging news stories," in *Proceedings of the Events and Stories in the News Workshop*. Vancouver, Canada: Association for Computational Linguistics, 2017. [Online]. Available: https://aclanthology.org/W17-2701 pp. 1–9.

[288] S. Miranda, A. Znotiņš, S. B. Cohen, and G. Barzdins, "Multilingual clustering of streaming news," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018. [Online]. Available: https://aclanthology.org/D18-1483 pp. 4535–4544.

[289] T. Staykovski, A. Barrón-Cedeno, G. Da San Martino, and P. Nakov, "Dense vs. sparse representations for news stream clustering." in *Text2Story Workshop at ECIR*, 2019, pp. 47–52.

[290] K. K. Saravanakumar, M. Ballesteros, M. K. Chandrasekaran, and K. McKeown, "Event-driven news stream clustering using entity-aware contextual embeddings," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.eacl-main.198 pp. 2330–2340.

[291] C. Kedzie, K. McKeown, and F. Diaz, "Predicting salient updates for disaster summarization," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, 2015. [Online]. Available: https://aclanthology.org/P15-1155 pp. 1608–1617.

[292] C. Kedzie, K. McKeown, and H. Daumé III, "Content selection in deep learning models of summarization," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018. [Online]. Available: https://aclanthology.org/D18-1208 pp. 1818–1828.

[293] L. Wang, C. Cardie, and G. Marchetti, "Socially-informed timeline generation for complex events," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, 2015. [Online]. Available: https://aclanthology.org/N15-1112 pp. 1055–1065.

[294] G. Binh Tran, M. Alrifai, and D. Quoc Nguyen, "Predicting relevant news events for timeline summaries," in *Proceedings of the 22nd International Conference on World Wide Web*, 2013. [Online]. Available: https://dl.acm.org/doi/10.1145/2487788.2487829 pp. 91–92.

[295] X. Chen, Z. Chan, S. Gao, M. Yu, D. Zhao, and R. Yan, "Learning towards abstractive timeline summarization," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, S. Kraus, Ed. ijcai.org, 2019. [Online]. Available: https://doi.org/10.24963/ijcai.2019/686 pp. 4939–4945.

[296] K.-H. Nguyen, X. Tannier, and V. Moriceau, "Ranking multidocument event descriptions for building thematic timelines," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014. [Online]. Available: https://aclanthology.org/C14-1114 pp. 1208–1217.

[297] Y. Liu and M. Lapata, "Hierarchical transformers for multi-document summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/P19-1500 pp. 5070–5081.

[298] A. Fabbri, I. Li, T. She, S. Li, and D. Radev, "Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/P19-1102 pp. 1074–1084.

[299] H. L. Chieu and Y. K. Lee, "Query based event extraction along a timeline," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004. [Online]. Available: https://dl.acm.org/doi/10.1145/1008992.1009065 pp. 425–432.

[300] R. Yan, L. Kong, C. Huang, X. Wan, X. Li, and Y. Zhang, "Timeline generation through evolutionary trans-temporal summarization," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, 2011. [Online]. Available: https://aclanthology.org/D11-1040 pp. 433–443.

[301] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang, "Evolutionary timeline summarization: a balanced optimization framework via iterative substitution," in *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, W. Ma, J. Nie, R. Baeza-Yates, T. Chua, and W. B. Croft, Eds. ACM, 2011. [Online]. Available: https://doi.org/10.1145/2009916.2010016 pp. 745–754.

[302] G. B. Tran, T. A. Tran, N.-K. Tran, M. Alrifai, and N. Kanhabua, "Leveraging learning to rank in an optimization framework for timeline summarization," in *SIGIR 2013 Workshop on Time-aware Information Access (TAIA)*, 2013. [Online]. Available: https://www.researchgate.net/publication/327945108_Leveraging_Learning_To_Rank_in_an_Optimization_Framework_for_Timeline_Summarization

[303] G. Tran, M. Alrifai, and E. Herder, "Timeline summarization from relevant headlines," in *European Conference on Information Retrieval*. Springer, 2015. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-16354-3_26 pp. 245–256.

[304] W. Y. Wang, Y. Mehdad, D. R. Radev, and A. Stent, "A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, 2016. [Online]. Available: https://aclanthology.org/N16-1008 pp. 58–68.

[305] S. Martschat and K. Markert, "A temporally sensitive submodularity framework for timeline summarization," in *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Brussels, Belgium: Association for Computational Linguistics, 2018. [Online]. Available: https://aclanthology.org/K18-1023 pp. 230–240.

[306] J. Steen and K. Markert, "Abstractive timeline summarization," in *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Hong Kong, China: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/D19-5403 pp. 21–31.

[307] J. Ansah, L. Liu, W. Kang, S. Kwashie, J. Li, and J. Li, "A graph is worth a thousand words: Telling event stories using timeline summarization graphs," in *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, and L. Zia, Eds. ACM, 2019. [Online]. Available: https://doi.org/10.1145/3308558.3313396 pp. 2565–2571.

[308] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of artificial intelligence research*, vol. 22, pp. 457–479, 2004. [Online]. Available: https://dl.acm.org/doi/10.5555/1622487.1622501

[309] A. Haghighi and L. Vanderwende, "Exploring content models for multi-document summarization," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado: Association for Computational Linguistics, 2009. [Online]. Available: https://aclanthology.org/N09-1041 pp. 362–370.

[310] M. Yasunaga, R. Zhang, K. Meelu, A. Pareek, K. Srinivasan, and D. Radev, "Graph-based neural multi-document summarization," in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Vancouver, Canada: Association for Computational Linguistics, 2017. [Online]. Available: https://aclanthology.org/K17-1045 pp. 452–462.

[311] A. Fabbri, I. Li, T. She, S. Li, and D. Radev, "Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/P19-1102 pp. 1074–1084.

[312] R. Barzilay, K. R. McKeown, and M. Elhadad, "Information fusion in the context of multi-document summarization," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College Park, Maryland, USA: Association for Computational Linguistics, 1999. [Online]. Available: https://aclanthology.org/P99-1071 pp. 550–557.

[313] K. Ganesan, C. Zhai, and J. Han, "Opinosis: A graph based approach to abstractive summarization of highly redundant opinions," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Beijing, China: Coling 2010 Organizing Committee, 2010. [Online]. Available: https://aclanthology.org/C10-1039 pp. 340–348.

[314] S. Banerjee, P. Mitra, and K. Sugiyama, "Multi-document abstractive summarization using ILP based multi-sentence compression," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, Q. Yang and M. J. Wooldridge, Eds. AAAI Press, 2015. [Online]. Available: http://ijcai.org/Abstract/15/174 pp. 1208–1214.

[315] L. Huang, L. Wu, and L. Wang, "Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.acl-main.457 pp. 5094–5107.

[316] L. Wu, Y. Chen, K. Shen, X. Guo, H. Gao, S. Li, J. Pei, and B. Long, "Graph neural networks for natural language processing: A survey," *ArXiv preprint*, vol. abs/2106.06090, 2021. [Online]. Available: https://arxiv.org/abs/2106.06090

[317] N. De Cao, W. Aziz, and I. Titov, "Question answering by reasoning across documents with graph convolutional networks," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/N19-1240 pp. 2306–2317.

[318] M. Ding, C. Zhou, Q. Chen, H. Yang, and J. Tang, "Cognitive graph for multi-hop reading comprehension at scale," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/P19-1259 pp. 2694–2703.

[319] A. Asai, K. Hashimoto, H. Hajishirzi, R. Socher, and C. Xiong, "Learning to retrieve reasoning paths over wikipedia graph for question answering," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=SJgVHkrYDH

[320] S. Min, D. Chen, L. Zettlemoyer, and H. Hajishirzi, "Knowledge guided text retrieval and reading for open domain question answering," *ArXiv preprint*, vol. abs/1911.03868, 2019. [Online]. Available: https://arxiv.org/abs/1911.03868

[321] R. Das, A. Godbole, D. Kavarthapu, Z. Gong, A. Singhal, M. Yu, X. Guo, T. Gao, H. Zamani, M. Zaheer, and A. McCallum, "Multi-step entity-centric information retrieval for multi-hop question answering," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Hong Kong, China: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/D19-5816 pp. 113–118.

[322] M. Li, Q. Zeng, Y. Lin, K. Cho, H. Ji, J. May, N. Chambers, and C. Voss, "Connecting the dots: Event graph schema induction with path language modeling," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.emnlp-main.50 pp. 684–695.

[323] D. Phung, T. N. Nguyen, and T. H. Nguyen, "Hierarchical graph convolutional networks for jointly resolving cross-document coreference of entity and event mentions," in *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*. Mexico City, Mexico: Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.textgraphs-1.4 pp. 32–41.

[324] "VOA News," https://www.voanews.com/.

[325] G. Ye, Y. Li, H. Xu, D. Liu, and S.-F. Chang, "Eventnet: A large scale structured concept library for complex event detection in video," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 471–480.

[326] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[327] X. Pan, B. Zhang, J. May, J. Nothman, K. Knight, and H. Ji, "Cross-lingual name tagging and linking for 282 languages," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017. [Online]. Available: https://aclanthology.org/P17-1178 pp. 1946–1958.

[328] C. Walker, S. Strassel, J. Medero, and K. Maeda, "Ace 2005 multilingual training corpus," *Linguistic Data Consortium, Philadelphia*, vol. 57, 2006.

[329] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider, "Abstract Meaning Representation for sembanking," in *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Sofia, Bulgaria: Association for Computational Linguistics, 2013. [Online]. Available: https://aclanthology.org/W13-2322 pp. 178–186.

[330] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl

[331] G. A. Miller, "WordNet: A lexical database for English," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992. [Online]. Available: https://aclanthology.org/H92-1116

[332] C. J. Fillmore, C. R. Johnson, and M. R. Petruck, "Background to framenet," *International journal of lexicography*, vol. 16, no. 3, pp. 235–250, 2003.

[333] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, "Zero-shot object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 384–400.

[334] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov et al., "The open images dataset v4," *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1956–1981, 2020.

[335] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland: Association for Computational Linguistics, 2014. [Online]. Available: https://aclanthology.org/P14-5010 pp. 55–60.

[336] L. Floridi and M. Chiriatti, "Gpt-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, no. 4, pp. 681–694, 2020.

[337] R. Sinkhorn, "A relationship between arbitrary positive matrices and doubly stochastic matrices," *The annals of mathematical statistics*, vol. 35, no. 2, pp. 876–879, 1964. [Online]. Available: https://projecteuclid.org/euclid.aoms/1177703591

[338] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, Eds., 2013. [Online]. Available: https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html pp. 2292–2300.

[339] H. Xu, D. Luo, and L. Carin, "Scalable gromov-wasserstein learning for graph partitioning and matching," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/6e62a992c676f611616097dbea8ea030-Abstract.html pp. 3046–3056.

[340] H. Shi, J. Mao, T. Xiao, Y. Jiang, and J. Sun, "Learning visually-grounded semantics from contrastive adversarial samples," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. [Online]. Available: https://aclanthology.org/C18-1315 pp. 3715–3727.

[341] J. S. Park, C. Bhagavatula, R. Mottaghi, A. Farhadi, and Y. Choi, "Visualcomet: Reasoning about the dynamic context of a still image," in *European Conference on Computer Vision*. Springer, 2020, pp. 508–524.

[342] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html pp. 5754–5764.

[343] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html pp. 5998–6008.

[344] J. Gu, Z. Lu, H. Li, and V. O. Li, "Incorporating copying mechanism in sequence-to-sequence learning," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016. [Online]. Available: https://aclanthology.org/P16-1154 pp. 1631–1640.

[345] Y. Lin, H. Ji, F. Huang, and L. Wu, "A joint end-to-end neural model for information extraction with global features," in *Proceedings of the 2020 Annual Meeting of the Association for Computational Linguistics (ACL2020)*, 2020.

[346] P. Budzianowski and I. Vulić, "Hello, it's GPT-2 - how can I help you? towards the use of pretrained language models for task-oriented dialogue systems," in *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Hong Kong: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/D19-5602 pp. 15–22.

[347] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray et al., "Training language models to follow instructions with human feedback," *ArXiv preprint*, vol. abs/2203.02155, 2022. [Online]. Available: https://arxiv.org/abs/2203.02155

[348] E. Filatova and V. Hatzivassiloglou, "Event-based extractive summarization," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004. [Online]. Available: https://aclanthology.org/W04-1017 pp. 104–111.

[349] X. Pan, T. Cassidy, U. Hermjakob, H. Ji, and K. Knight, "Unsupervised entity linking with Abstract Meaning Representation," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, 2015. [Online]. Available: https://aclanthology.org/N15-1119 pp. 1130–1139.

[350] T. Lai, H. Ji, T. Bui, Q. H. Tran, F. Dernoncourt, and W. Chang, "A context-dependent gated module for incorporating symbolic semantics into event coreference resolution," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.naacl-main.274 pp. 3491–3499.

[351] Q. Ning, S. Subramanian, and D. Roth, "An improved neural baseline for temporal relation extraction," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/D19-1642 pp. 6203–6209.

[352] H. Wen, Y. Qu, H. Ji, Q. Ning, J. Han, A. Sil, H. Tong, and D. Roth, "Event time extraction and propagation via graph attention networks," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021. [Online]. Available: https://aclanthology.org/2021.naacl-main.6 pp. 62–73.

[353] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/N19-1423 pp. 4171–4186.

[354] T. Ma and J. Chen, "Unsupervised learning of graph hierarchical abstractions with differentiable coarsening and optimal transport," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17072 pp. 8856–8864.

[355] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl

[356] U. Manber, *Introduction to algorithms: a creative approach*. Addison-Wesley Longman Publishing Co., Inc., 1989. [Online]. Available: https://dl.acm.org/doi/10.5555/534662

[357] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford et al., "Okapi at trec-3," *Nist Special Publication Sp*, vol. 109, p. 109, 1995. [Online]. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/okapi_trec3.pdf

[358] M. Li, A. Zareian, Q. Zeng, S. Whitehead, D. Lu, H. Ji, and S.-F. Chang, "Cross-media structured common space for multimedia event extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. [Online]. Available: https://aclanthology.org/2020.acl-main.230 pp. 2557–2568.

[359] H. Zheng and M. Lapata, "Sentence centrality revisited for unsupervised summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. [Online]. Available: https://aclanthology.org/P19-1628 pp. 6236–6247.

[360] J. Zhao, M. Liu, L. Gao, Y. Jin, L. Du, H. Zhao, H. Zhang, and G. Haffari, "Summpip: Unsupervised multi-document summarization with sentence graph compression," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, J. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, and Y. Liu, Eds. ACM, 2020. [Online]. Available: https://doi.org/10.1145/3397271.3401327 pp. 1949–1952.

[361] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[362] H. He, H. Zhang, and D. Roth, "Rethinking with retrieval: Faithful large language model inference," *ArXiv preprint*, vol. abs/2301.00303, 2023. [Online]. Available: https://arxiv.org/abs/2301.00303

[363] Z. Yu, C. Xiong, S. Yu, and Z. Liu, "Augmentation-adapted retriever improves generalization of language models as generic plug-in," *ArXiv preprint*, vol. abs/2305.17331, 2023. [Online]. Available: https://arxiv.org/abs/2305.17331

[364] W. Yu, Z. Zhang, Z. Liang, M. Jiang, and A. Sabharwal, "Improving language models via plug-and-play retrieval feedback," *ArXiv preprint*, vol. abs/2305.14002, 2023. [Online]. Available: https://arxiv.org/abs/2305.14002

[365] K.-H. Huang, H. P. Chan, and H. Ji, "Zero-shot faithful factual error correction," *ArXiv preprint*, vol. abs/2305.07982, 2023. [Online]. Available: https://arxiv.org/abs/2305.07982

[366] S. Li, M. Namazifar, D. Jin, M. Bansal, H. Ji, Y. Liu, and D. Hakkani-Tur, "Enhancing knowledge selection for grounded dialogues via document semantic graphs," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, 2022. [Online]. Available: https://aclanthology.org/2022.naacl-main.202 pp. 2810–2823.

[367] T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling large language models to generate text with citations," *ArXiv preprint*, vol. abs/2305.14627, 2023. [Online]. Available: https://arxiv.org/abs/2305.14627

[368] O. Weller, M. Marone, N. Weir, D. Lawrie, D. Khashabi, and B. Van Durme, "" according to..." prompting language models improves quoting from pre-training data," *ArXiv preprint*, vol. abs/2305.13252, 2023. [Online]. Available: https://arxiv.org/abs/2305.13252

[369] F. Petroni, P. Lewis, A. Piktus, T. Rocktäschel, Y. Wu, A. H. Miller, and S. Riedel, "How context affects language models' factual predictions," *ArXiv preprint*, vol. abs/2005.04611, 2020. [Online]. Available: https://arxiv.org/abs/2005.04611

[370] M. Geva, J. Bastings, K. Filippova, and A. Globerson, "Dissecting recall of factual associations in auto-regressive language models," *ArXiv preprint*, vol. abs/2304.14767, 2023. [Online]. Available: https://arxiv.org/abs/2304.14767

[371] D. Dai, L. Dong, Y. Hao, Z. Sui, B. Chang, and F. Wei, "Knowledge neurons in pretrained transformers," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022. [Online]. Available: https://aclanthology.org/2022.acl-long.581 pp. 8493–8502.

[372] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and editing factual associations in gpt," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 359–17 372, 2022.

[373] K. Meng, A. S. Sharma, A. Andonian, Y. Belinkov, and D. Bau, "Mass-editing memory in a transformer," *ArXiv preprint*, vol. abs/2210.07229, 2022. [Online]. Available: https://arxiv.org/abs/2210.07229

[374] N. Lee, W. Ping, P. Xu, M. Patwary, P. N. Fung, M. Shoeybi, and B. Catanzaro, "Factuality enhanced language models for open-ended text generation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34 586–34 599, 2022.

[375] A. Chen, P. Pasupat, S. Singh, H. Lee, and K. Guu, "Purr: Efficiently editing language model hallucinations by denoising language model corruptions," *ArXiv preprint*, vol. abs/2305.14908, 2023. [Online]. Available: https://arxiv.org/abs/2305.14908

[376] J. Wei, L. Hou, A. Lampinen, X. Chen, D. Huang, Y. Tay, X. Chen, Y. Lu, D. Zhou, T. Ma et al., "Symbol tuning improves in-context learning in language models," *ArXiv preprint*, vol. abs/2305.08298, 2023. [Online]. Available: https://arxiv.org/abs/2305.08298

[377] K. Yang, C. Yu, Y. Fung, M. Li, and H. Ji, "Adept: A debiasing prompt framework," *Proc. Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI2023)*, 2023.

[378] C. Yu, S. Jeoung, A. Kasi, P. Yu, and H. Ji, "Unlearning bias in language models by partitioning gradients," in *Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023) Findings*, 2023.

[379] A. Omrani, A. S. Ziabari, C. Yu, P. Golazizian, B. Kennedy, M. Atari, H. Ji, and M. Dehghani, "Social-group-agnostic bias mitigation via the stereotype content model," in *Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023)*, 2023.