Research Statement

Manling Li

My research vision is to equip machines with **deep semantic understandings of multimodal information**. *What happened? Who? When? Where? Why? What will happen next?* are the fundamental questions asked to comprehend the overwhelming amount of information available. 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, I focus 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 my 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, 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.

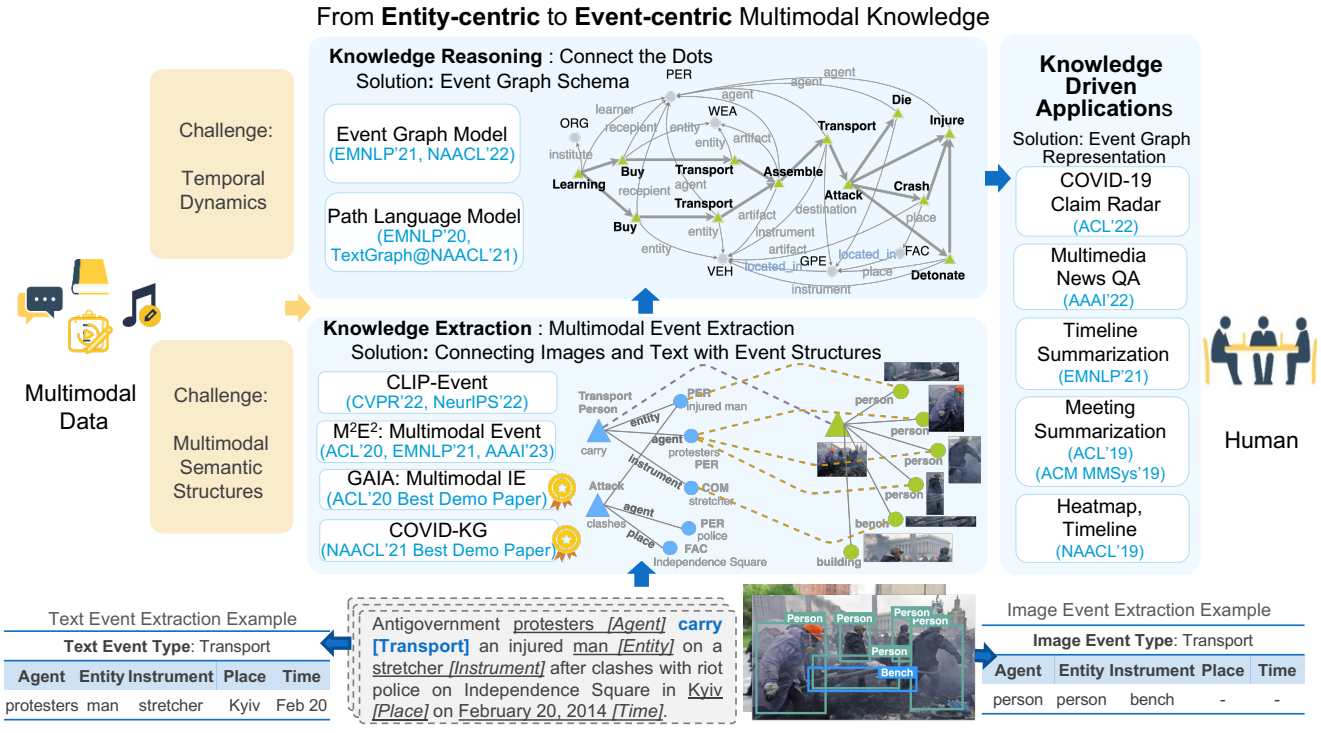

Figure 1: My dissertation is driven by the *multimodal event semantic structure* and *temporal dynamics* challenges. Our novel *Multimodal Event Graph* structural knowledge representation bridges event knowledge across multiple modalities using a unified structure that is machine-readable.

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. As the first to **introduce complex event semantic structures into vision-language pretraining** (CLIP-Event) [13, 19]), I propose a **zero-shot cross-modal transfer** of semantic understanding abilities from language to vision, which resolves the poor portability issue and supports **Zero-shot Multimodal Event Extraction** (M^2E^2) [15, 2] for the first time. I led the development of open-source Multimodal IE system GAIA [14, 8], which was awarded **ACL 2020 Best Demo Paper**, and **ranked 1st** in DARPA AIDA evaluation each phase. Our COVID knowledge graph (COVID-KG) [18] was awarded **NAACL 2021 Best Demo Paper** and the released knowledge graph has been widely used by other researchers (downloaded more than 2000 times since 2021).
- 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. I propose **Event Graph Schema** [16, 5, 3], which open doors to a global event graph context to enable alternative predictions, along with structural justifications including location-, attribute-, and participant-specific details. My event tracking system [7] was selected to be presented at the **ARL NS-CTA project 10-year Milestone Demo Day** in 2019 and at the **DARPA Demo Day** in 2019.
- Applying Event Knowledge Graph (Knowledge Driven Applications): My work has shown positive results on long-standing open problems, such as Timeline Summarization [11, 10], Meeting Summarization [17], and Multimedia News Question Answering [4]. But the impact of my work extends beyond academia — one of my works [10] was **patented during collaborations with IBM Research**, and multiple systems that I led have been **transferred to DARPA and ARL** for intelligent analysis, including Cross-media Cross-lingual News Recommendation [14] and COVID-19 Claim Radar [12].

My work on Multimedia Event Knowledge Graphs opens doors to the **next generation of information access** to understand abstract events that are multimodal, inter-connected, and predictive of arguments. My work has been recognized by industry (**Microsoft Research PhD Fellowship**), by academia (**EE CS Rising Star**), and by the government (**DARPA Riser** with a presentation at the **DARPA Forward Event**). Our tutorials on *Event-Centric Natural Language Processing* [1] attracted around 200 conference members at ACL 2021 and AAAI 2021.

Research Philosophy. It is my firm belief that the purpose of AI is not to boost scores on specific tasks, but to aid human capabilities in real world, such as in reading, seeing, hearing and making decisions with far-reaching consequences. I am committed to grounding such techniques in a social good context for healthier information consumption and dissemination, such as utilizing multimedia (languages, visual aids) to communicate knowledge and to present trustworthy and explainable knowledge for humans.

Understanding Multimodal Semantic Structures for Knowledge Extraction

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. A comprehensive understanding of events requires computers to perform joint comprehension across multiple modalities, such as text, images, and videos.

My work is founded upon **a brand new research direction, Multimedia Event Extraction (M^2E^2)** [15], 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).

My solution to joint event structure extraction is to construct a **multimodal common semantic space via Vision-Language (V+L) pretraining that preserves event semantic structures**, i.e., similar events and their arguments are close in this embedding space regardless of their source modality. I propose **CLIP-Event** [13] and **VidIL** [19] to **transfer such event knowledge from text to images in a zero-shot manner**. My 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 19 students** to develop and release **the first open-source multimedia knowledge extraction system GAIA** [14, 7, 9, 8]. 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**. The effectiveness extend to scientific literature. We released a multimedia **Scientific Information Extraction system COVID-KG** [18], 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**.

Understanding Temporal Dynamics for Knowledge Reasoning

Event extraction from massive multimedia data enables us to obtain a large number of historical events. These historical events imply knowledge about event interactions, which guides our predictions as to what might happen next and what events are missing. For example, after an ATTACK event there will usually be ARREST and SENTENCE 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. However, previous event schema induction work has been oversimplified to be local and sequential, rendering them incapable of dealing with real-world events with numerous participants, complicated timelines, and various alternative outcomes.

My research tackles this issue with a new paradigm of event schema knowledge: an **Event Graph Schema** [16, 5, 3], which is a graph-based schema representation that encompasses events, arguments, temporal connections and argument relations. **Figure 1 shows an example schema** of the complex event type *car-bombing*: a person learns to make bombs, purchasing materials and a vehicle; a bomb is fixed to the vehicle; and the attacker drives the vehicle to attack people. In this scenario, people can be hurt by the vehicle or by the bomb's explosion. It is the first application of graph generation to induce event schemas and predict future events.

My work is **a new step towards the semantic understanding of inter-event connections**. Different from traditional methods using one-hop relations as connections between events, I 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, my 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.

Knowledge Driven Applications

One of the main bottlenecks of large corpora analysis is the multi-document encoding. 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).

My solution is using event graphs to provide a new comprehensive representation and necessary inductive bias. I 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. These event graphs can then be used to address the massive unstructured data challenge in real-world applications: (1) **Timeline Summarization** [11, 10] is formulated as an event graph compression problem and then I design time-aware optimal transport to obtain the summary graph. (2) **Meeting Summarization** [17] 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** [4] employs multimedia event graphs to condition synthetic question-answer generation, and to automatically augment data via weak supervision.

My work 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.

Future Research Agenda

My **long-term goal** is to make machines capable of understanding deep semantics as humans do, especially abstract semantics, including semantic roles of objects (such as *victim*, *detainee*), and semantics of abstract concepts (such as *love*, *happiness*). **This structured view of knowledge enables machines to further comprehend, reason, and communicate knowledge through vision and natural language.** I plan to continue my research along this path by approaching from the following directions:

Abstract Semantics Understanding via Compositional Cross-modal Transfer Learning. I aim to build open-world vision-language and reasoning models that reason about abstract concepts, from general, simple, and observational to specific, complex, and interpretive. I aim to propose a neural symbolic reasoning framework that is able to compositionally learn new concepts with the training signal transferred from language. 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. I am passionate about working with excellent 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.

Socially-Minded Healthier Information Consumption. 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. I am highly interested

in identifying writers' opinions, intentions and hidden agendas, thereby reducing ambiguity in text by revealing any underlying meaning and bias. I am interested in collaborating with researchers in **Psychology, Computational Social Science, and Data Mining** to automatically discern participants' demographic characteristics, region-level factors, geographical indicators, cultural backgrounds, and social values, which in turn could lead to improved safety and societal justice.

Structured Knowledge-Driven Explainable and Trustworthy AI. Structured knowledge is effective to extract salient information and convert large-scale corpora into easily queryable formats that allow for tracking, cross-checking and verification. I aim to force the model to provide rationales for its output decisions, and allow the model to better align its internal decision-making process to the set of key decision-maker knowledge in both structured and unstructured formats. I am excited to work with researchers in **Human Computer Interaction and Symbolic AI** to improve human-machine collaboration performance as a whole, and further move forward trustworthiness and explainability of AI models.

Interacting with the Physical World. The grounding of event and entity knowledge in the real world requires physical knowledge, such as affordance, visual state changes, location information, etc. I am passionate about predicting future events through both physical properties and the potential interactions between an object and the environment. I would be glad to collaborate with researchers in **Robotics and Embodied AI** to improve language models by integrating them into physical environments, as well as to incorporate language signals into autonomous driving and task planning.

Collaboration and Funding

The research directions that I intend to explore require collaboration with expert researchers in many fields, including natural language processing, computer vision, machine learning, data mining, social science, robotics, human-machine interaction, systems and algorithms. I had rich experiences to lead and manage large projects involving more than 20 people and coordinate with multiple universities and institutes. **I led a 19-student team** to develop the UIUC Knowledge Extraction System, and coordinated with 4 professors from different universities to participate in the DARPA AIDA Evaluations of Phase I, II and III. **I am fortunate to have close collaborations with 21 professors from 15 universities and research institutes**, such as Columbia University, New York University (NYU), University of Pennsylvania (UPenn), University of Southern California (USC), University of North Carolina at Chapel Hill (UNC), Florida University, etc. **I also have had the fortune to work closely with researchers in fields outside of computer science**, including psychology, biology, and healthcare, to conduct interdisciplinary research. I plan to continue existing collaborations and foster new connections in order to develop well-established principles underlying multimodal knowledge research.

During my PhD, my work has been mainly supported by the Microsoft Research PhD Fellowship, DARPA, ARL, ARFL, IARPA, and NSF. I have been nominated by DARPA as a **DARPA Riser** (awarded to young faculties, postdocs, and senior PhDs who will lead to technological breakthroughs), and I presented ideas directly to DARPA project managers during the invitation-only DARPA Forward Event. Additionally, **I have contributed significantly to the writing of winning grant proposals**, including idea generation, method design, idea illustration and visual aid creation, such as DARPA KAIROS project, DARPA ITM project, DARPA SemaFor project, DARPA CCU project, and NSF MMLI project. I will continue to seek funding opportunities in the future from multiple funding agencies (e.g., DARPA, ARL, ARFL, IARPA, NSF, NIH) and industries.

References

[1] Muhao Chen, Hongming Zhang, Qiang Ning, **Manling Li**, Heng Ji, Kathleen McKeown, and Dan Roth. Event-centric natural language processing. *ACL Tutorial*, 2021.

[2] Yang Guang, **Manling Li**, Jiajie Zhang, Xudong Lin, Shih-Fu Chang, and Heng Ji. Video event extraction via tracking visual states of arguments. *EMNLP Findings*, 2021.

[3] Xiaomeng Jin, **Manling Li**, and Heng Ji. Event schema induction with double graph autoencoders. *NAACL*, 2022.

[4] Revanth Reddy, Xilin Rui, **Manling Li**, Xudong Lin, Haoyang Wen, Jaemin Cho, Lifu Huang, Mohit Bansal, Avi Sil, Shih-Fu Chang, Alexander Schwing, and Heng Ji. Multi-media multi-hop news question answering via cross-media grounding. *AAAI*, 2021.

[5] **Manling Li**, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, and Jiawei Han. The future is not one-dimensional: Graph modeling based complex event schema induction for event prediction. *EMNLP*, 2021.

[6] **Manling Li**, Xudong Lin, Jie Lei, Mohit Bansal, Heng Ji, and Shih-Fu Chang. Knowledge-driven vision-language pretraining. *AAAI Tutorial*, 2023.

[7] **Manling Li**, Ying Lin, Joseph Hoover, Spencer Whitehead, Clare Voss, Morteza Dehghani, and Heng Ji. Multilingual entity, relation, event and human value extraction. In *NAACL (Demonstrations)*, 2019.

[8] **Manling Li**, Ying Lin, Tuan Manh Lai, Xiaoman Pan, Haoyang Wen, Sha Li, Zhenhailong Wang, Pengfei Yu, Lifu Huang, Di Lu, Qingyun Wang, Haoran Zhang, Qi Zeng, Chi Han, Zixuan Zhang, Yujia Qin, Xiaodan Hu, Nikolaus Parulian, Daniel Campos, Heng Ji, Brian Chen, Xudong Lin, Alireza Zareian, Amith Ananthram, Emily Allaway, Shih-Fu Chang, Kathleen McKeown, Yixiang Yao, Yifan Wang, Michael Spector, Mitchell DeHaven, Daniel Napierski, Marjorie Freedman, Pedro Szekely, Haidong Zhu, Ram Nevatia, Yang Bai, Yifan Wang, Ali Sadeghian, Haodi Ma, and Daisy Zhe Wang. Gaia at sm-kbp 2020 - a dockerized multi-media multi-lingual knowledge extraction, clustering, temporal tracking and hypothesis generation system. In *NIST TAC KBP*, 2020.

[9] **Manling Li**, Ying Lin, Ananya Subburathinam, Spencer Whitehead, Xiaoman Pan, Di Lu, Qingyun Wang, Tongtao Zhang, Lifu Huang, Heng Ji, Alireza Zareian, Hassan Akbari, Brian Chen, Bo Wu, Emily Allaway, Shih-Fu Chang, Kathleen McKeown, Yixiang Yao, Jennifer Chen, Eric Berquist, Kexuan Sun, Xujun Peng, Ryan Gabbard, Marjorie Freedman, Pedro Szekely, T.K. Satish Kumar, Arka Sadhu, Ram Nevatia, Miguel Rodriguez, Yifan Wang, Yang Bai, Ali Sadeghian, and Daisy Zhe Wang. Gaia at sm-kbp 2019-a multi-media multi-lingual knowledge extraction and hypothesis generation system. In *NIST TAC KBP*, 2019.

[10] **Manling Li**, Tengfei Ma, Mo Yu, and Achille Fokoue. Unsupervised knowledge graph compression based on optimal transport. *U.S. Patent submission*, 2020.

[11] **Manling Li**, Tengfei Ma, Mo Yu, Lingfei Wu, Tian Guo, Heng Ji, and Kathleen McKeown. Timeline summarization based on event graph compression via time-aware optimal transport. *EMNLP*, 2021.

[12] **Manling Li**, Revanth Reddy, Haoyang Wen, and Heng Ji. Covid-19 claim radar: A structured claim extraction and tracking system. In *ACL (Demonstrations)*, 2022.

[13] **Manling Li**, Ruochen Xu, Shuohang Wang, Xudong Lin, Chenguang Zhu, Xuedong Huang, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting vision and text with event structures. *CVPR*, 2022.

[14] **Manling Li**, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare R Voss, Dan Napierski, and Marjorie Freedman. Gaia: A fine-grained multimedia knowledge extraction system. In *ACL (Demonstrations)*, 2020.

[15] **Manling Li**, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. Cross-media structured common space for multimedia event extraction. In *ACL*, 2020.

[16] **Manling Li**, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, May, Jonathan, Nathanael Chambers, and Clare Voss. Connecting the dots: Event graph schema induction with path language modeling. In *EMNLP*, 2020.

[17] **Manling Li**, Lingyu Zhang, Heng Ji, and Richard J Radke. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *ACL*, 2019.

[18] Qingyun Wang, **Manling Li**, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Haoran Zhang, Weili Liu, Aabhas Chauhan, Yingjun Guan, Bangzheng Li, Ruisong Li, Xiangchen Song, Heng Ji, Jiawei Han, Shih-Fu Chang, James Pustejovsky, David Liem, Ahmed Elsayed, Martha Palmer, Jasmine Rah, Clare Voss, Cynthia Schneider, and Boyan Onyshkevych. Covid-19 literature knowledge graph construction and drug repurposing report generation. In *NAACL (Demonstrations)*, 2021.

[19] Zhenhailong Wang, **Manling Li**, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, Shih-Fu Chang, Mohit Bansal, and Heng Ji. Language models with image descriptors are strong few-shot video-language learners. *NeurIPS*, 2022 (equal contribution).