LEARNING VERSATILE MULTIMODAL REPRESENTATION FOR KNOWLEDGE EXTRACTION AND REASONING

CHANGMENG ZHENG

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Computing

Learning Versatile Multimodal Representation for Knowledge Extraction and Reasoning

Changmeng Zheng

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Jun 2024

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Signature: ____________________

Name of Student: Changmeng Zheng

Abstract

Relational facts organize human knowledge of the real world in a triplet format. These structural facts are regarded as the way to implement conscious and logical intelligence. Although the past three decades have witnessed the rise of text analysis methods for extracting meaningful information from unstructured textual data, these methods often fall short of capturing the full semantic richness and complexity of human language, particularly when it comes to understanding the relationships between entities. Besides, textual semantics are sometimes incomplete and ambiguous, which can cause inaccuracy and severe misleading in facts. On the contrary, the information from other modalities (e.g., visual contents) is much more intuitive and specific. Inspired by the human capacity to perceive and communicate through a multisensory system, this thesis explores the potential of learning versatile multimodal representations for knowledge extraction and reasoning. This thesis delves into four critical challenges within multimodal learning, proposing novel solutions through a series of rigorous investigations:

(1) A Unified Multimodal Graph Learning Framework: To overcome the prevalent issues of modality gaps and spurious alignments in multimodal knowledge extraction, we present a novel multimodal graph learning framework. This framework enables a comprehensive mapping of diverse elements from disparate modalities onto a unified graph structure. By emphasizing the capture of fine-grained correlations through semantic and structural graph alignment, we achieve improved knowledge

extraction accuracy. Additionally, we introduce a benchmark dataset specifically designed for this task, empirically validating the efficacy of our proposed framework.

(2) A Hierarchical Multimodal Representation Learning Method: To address the limitations of inconsistent semantic levels between individual modality representations, we further explore the integration of hierarchical multimodal learning by incorporating information at different granularities (e.g., from image-level to object-level visual features and from sentence-level to concept-level textual features). By connecting vision and language through paths within external concept graphs, we bridge the gap between modalities, mirroring the human association process.

(3) A Robust Data Augmentation and Estimation System: To acknowledge the detrimental impact of misalignment issue in text-image datasets, we investigate methods for mitigating bias and distractions caused by such misalignments. Drawing inspiration from machine translation techniques, this work employs back-translation and divergence estimation to identify and reduce the influence of irrelevant or partially aligned information, leading to more robust and reliable knowledge extraction.

(4) An Iterative Refined Graph Reasoning Application: To demonstrate the generality and versatility of the extracted multimodal knowledge graph, we incorporate multi-agent debate into multimodal reasoning to facilitate iterative refinement of knowledge representations. The proposed Blueprint Debate on Graphs framework utilizes a graph-based structure for representing and refining knowledge, encouraging collaboration and competition between agents to achieve a deeper understanding of the relationships and interactions within multimodal data.

By addressing the challenges of fine-grained alignment, hierarchical learning, bias mitigation, and iterative refinement, this research contributes to the advancement of multimodal learning across several tasks and benchmarks, and unlocks new possibilities for understanding and utilizing the rich information embedded within multimodal data.

Keywords: Multimodal Knowledge Extraction, Multimodal Reasoning, Graph Alignment, Hierarchical Representation Learning, Diffusion Models, Multi-agent Collaboration.

Publications Arising from the Thesis

1. Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li and Tao Wang, "Multimodal Relation Extraction with Efficient Graph Alignment", in *Proceedings of the 29th ACM international conference on multimedia (ACM MM)*, pp. 5298-5306, 2021. (**Chapter 3**)

2. Junhao Feng, Guohua Wang, Changmeng Zheng (Corresponding Author), Yi Cai, Ze Fu, Yaowei Wang, Xiao-Yong Wei and Qing Li, "Towards Bridged Vision and Language: Learning Cross-modal Knowledge Representation for Relation Extraction", in *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, volume 34, issue 1, pp. 561-575, 2023. (**Chapter 4**)

3. Changmeng Zheng, Junhao Feng, Yi Cai, Xiaoyong Wei and Qing Li, "Rethinking Multimodal Entity and Relation Extraction from a Translation Point of View", in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6810-6824, 2023. (**Chapter 5**)

4. Changmeng Zheng, Dayong Liang, Wengyu Zhang, Xiao-Yong Wei, Tat-Seng Chua, Qing Li, "A Picture Is Worth a Graph: A Blueprint Debate Paradigm for Multimodal Reasoning", in *Proceedings of the 32th ACM international conference on multimedia (ACM MM)*, pp. 419-428, 2024. (**Chapter 6**)

Acknowledgments

I would like to express my heartfelt gratitude to everyone who has supported me throughout my studies and contributed to the work presented in this thesis.

First and foremost, I extend my deepest thanks to my supervisor, Prof. Qing Li. His unwavering support and guidance have been instrumental in both my academic journey and personal growth at The Hong Kong Polytechnic University. Prof. Li has profoundly enhanced my critical thinking and taught me the nuances of conducting professional research. I am deeply appreciative of his generous contributions of ideas, time, and thoughtful consideration over the past three years. It has been an immense honor to be his Ph.D. student.

I am particularly grateful to Prof. Xiao-Yong Wei for his invaluable advice and contributions. Collaborating with him and the Peng Cheng Laboratory has allowed me to delve into the fascinating world of emerging Large Foundation Models and broaden my knowledge across various research domains. I would also like to extend my sincere thanks to my colleagues in our group: Zehang Lin, Da Ren, Jiatong Li, Yujuan Ding, Yaowei Wang, and Sirui Huang, for their thought-provoking discussions, joint efforts, and unwavering mutual support.

My gratitude also extends to Prof. Tat-Seng Chua, for his expert guidance and generous encouragement during my visit to the National University of Singapore. His insightful suggestions have led me to explore significant research problems in the field of multimodal Large Language Models and inspired me to seek feasible solutions to these challenges. It has been a privilege to learn from and work with Prof. Chua. I am equally thankful to every member of the NExT++ Lab at the National University of Singapore, including Yunshan Ma, Ao Zhang, Yang Deng, Xue Li, Jujia Zhao, An Zhang, Haoxuan Li, Jinghao Zhang, and Moxin Li. It has been an honor to collaborate with and befriend such a talented group of individuals.

Lastly, but most importantly, my deepest appreciation is reserved for my family, especially my wife, Heyang Peng. Her belief in me has been a constant source of inspiration and motivation. She has been with me through countless late nights and early mornings, always ready to offer a comforting word or a reassuring smile. Her sacrifices, understanding, and unwavering faith in my abilities have carried me through the toughest times. Thank you for being my rock, my confidante, and my greatest supporter.

I extend my gratitude to all who have been part of this journey. As the old saying goes, “It takes a village to raise a child.” In my case, it has taken a community of mentors, colleagues, friends, and family to help me reach this milestone. Your collective support has been invaluable, and I am deeply grateful for your presence in my life. Thank you all.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background: Multimodal Knowledge Extraction and Reasoning

Similar to the challenges face by users in the ever-expanding landscape of the World Wide Web, where information overload become a significant hurdle, the realm of data management and analysis also encounters a critical juncture [120]. The exponential growth of data, encompassing structured, semi-structured, and unstructured formats, demands innovative approaches to organize, understand, and utilize this vast information reservoir. Knowledge graphs [50] have emerged as a powerful solution to address this challenge, offering a means to represent and reason about knowledge in a more structured and interconnected manner. For example, one could derive the knowledge triplet (Joe Biden, president of, USA) from the sentence “Joe Biden serves as the 46th president of the United States”. While early knowledge representation systems like semantic networks and ontologies lay the groundwork, the concept of knowledge graphs gains significant traction with the introduction of Google’s Knowledge Graph in 2012. These large-scale knowledge graphs, composed of entities and concepts as nodes interconnected by diverse semantic relationships, have demonstrated immense value across a wide spectrum of real-world applications. Examples include enhancing text understanding in natural language processing, bolstering the effectiveness of recommendation systems [29], and powering sophisticated natural language question answering systems [73].

The majority of current knowledge extraction (KE) methods focus on extracting entities and their relationships from purely linguistic contexts, such as sentences, employing a discriminative approach. This means they typically classify entity types and relations from a predefined set of categories. While these methods have yielded significant progress [86, 82], they often overlook a crucial aspect of human cognition: our perception and interaction with the world occur through a multi-sensory system, encompassing not just language but also visual, auditory, and other sensory inputs. This limitation restricts the ability of existing KE methods to fully capture and understand the rich semantics underlying relationships. To illustrate, consider the concept of *parenting*. A human being's understanding of this action stems not solely from its textual definition but also from the lived experiences of being a parent or observing parental relationships. Similarly, without any prior visual reference, a person might misinterpret the photographer's request for a *hand-in-waistcoat* pose, lacking the visual understanding that it refers to a specific posture with the hand placed inside the coat flap.

On the other hand, it is important to acknowledge the domain-specific challenges that arise when applying these techniques to different types of data. Much of the existing work in KE focuses on the newswire domain, where language tends to be formal, complete, and structured. However, when dealing with user-generated content from social media platforms, these methods often encounter limitations due to the unique characteristics of such data [86]. For example, an image accompanying a social media post can provide visual cues that help disambiguate entity types or clarify the meaning of informal language. Consider a post mentioning "the GOAT,"

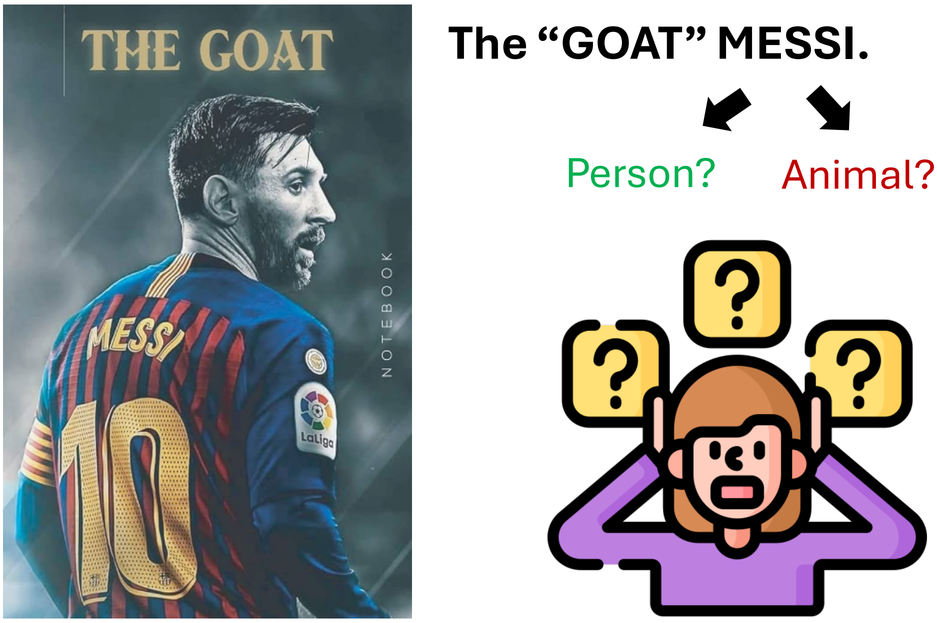


Figure 1.1: An example of multimodal named entity recognition (MNER) in social media posts from Twitter, where “MESSI” is the name of a Person instead of an Animal.

accompanied by a picture of Lionel Messi. The image clarifies that ”GOAT” refers to ”Greatest Of All Time” in the context of soccer and identifies Messi as the entity person in question. Similarly, video content can offer additional context and insights into events or activities mentioned in the text, enriching the understanding of the situation. By embracing a multi-modal approach, KE methods can overcome the limitations of domain-specific challenges and unlock the rich knowledge embedded within social media data. This not only expands the scope of knowledge extraction but also contributes to a more comprehensive and nuanced understanding of human communication and interaction in the digital age.

1.2 Motivations and Contributions

Deep Neural Networks (DNNs) have revolutionized various fields, including speech recognition [2], computer vision [37], and natural language processing [124], by demonstrating remarkable success in learning effective feature representations from complex data. More recently, the advent of attention mechanisms has further propelled the ability to model intricate correlations within and across different modalities, leading to the emergence of Attention-based Deep Neural Networks, such as transformers [22], as a dominant paradigm for multimodal representation learning [110], sequence modeling [116], and generation tasks [88].

Unsurprisingly, these powerful methods have also been adopted to address the challenges of multimodal knowledge extraction [16, 17]. However, existing approaches often fall short by treating visual and textual information separately, encoding them individually and then employing simple fusion techniques, such as addition or concatenation, to combine the resulting features. This neglects the inherent ”modality gap” – the fundamental differences in how information is represented and processed across different modalities – leading to suboptimal performance.

Therefore, there is a pressing need to develop a universally effective multimodal representation learning framework that can effectively harness the power of diverse modalities for knowledge extraction. Such a framework should go beyond shallow fusion techniques and delve deeper into understanding the interplay and interactions between modalities, capturing the rich semantic relationships that underpin knowledge representation.

This thesis aims to address the aforementioned challenges by proposing a novel framework for learning versatile multimodal representations specifically tailored for knowledge extraction and reasoning tasks. The core of this framework lies in a graph learning approach that effectively unifies diverse modalities through an efficient graph alignment strategy. Within this overarching framework, the thesis tackles three cru-

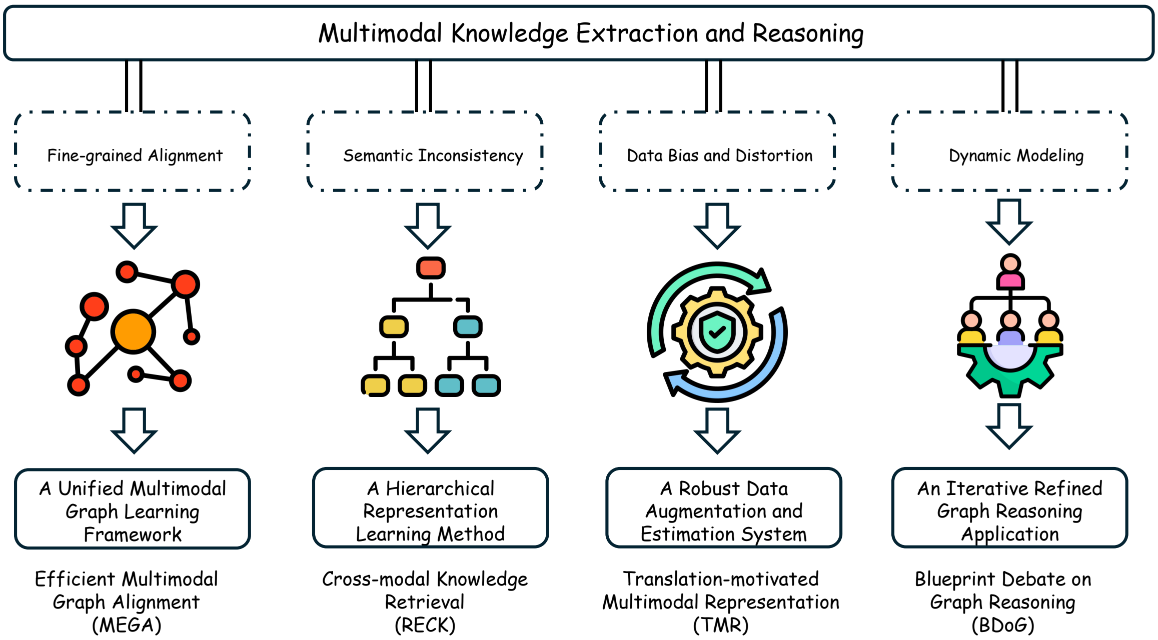

Multimodal Knowledge Extraction and Reasoning
Fine-grained Alignment
Semantic Inconsistency
Data Bias and Distortion
Dynamic Modeling
A Unified Multimodal Graph Learning Framework
A Hierarchical Representation Learning Method
A Robust Data Augmentation and Estimation System
An Iterative Refined Graph Reasoning Application
Efficient Multimodal Graph Alignment (MEGA)
Cross-modal Knowledge Retrieval (RECK)
Translation-motivated Multimodal Representation (TMR)
Blueprint Debate on Graph Reasoning (BDoG)

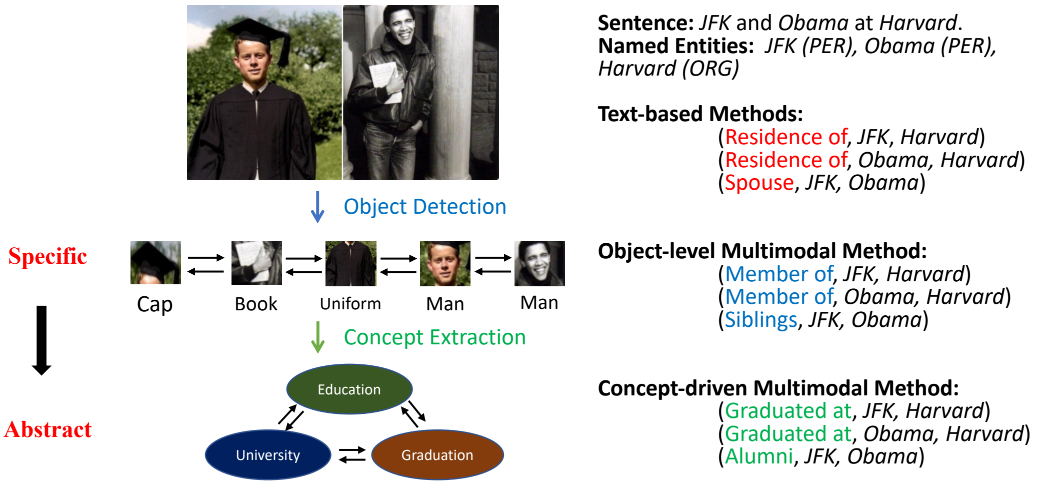
Figure 1.2: This thesis delves into four critical challenges within multimodal learning for knowledge extraction and reasoning.

cial issues inherent in multimodal learning: (1) Semantic Level Inconsistency: Bridging the gap between the inherently different semantic levels at which vision and language operate, ensuring a coherent and meaningful representation of information across modalities. (2) Bias and Distortion in Multimodal Alignment Data: Addressing potential biases and distortions present in multimodal alignment data, ensuring the robustness and reliability of the learned representations. (3) Iterative Refined Multimodal Graph and Its Application in Reasoning: Developing methods for dynamic updates of the multimodal graph, enabling adaptive learning and facilitating complex reasoning tasks involving multimodal knowledge. The following sections will delve deeper into each of these key challenges and present the proposed solutions for achieving robust and versatile multimodal representations for knowledge extraction and reasoning.

1.2.1 A Unified Multimodal Graph Learning Framework

The task of KE is to utilize text-related modality information (e.g., image posts) to supplement the missing semantics of short texts. In addition to the traditional challenges of inter-modal heterogenous gap, the task of KE faces more challenges that it need to extract the positive auxiliary knowledge from the possible noise alignment results. The semantic shifts of alignment of different modalities in the representations space further misleads the entity and relation inference in the text space. How to reduce the modality gap and obtain the knowledge from the noise alignment remains a challenging problem. Compared to those methods devoted to complicated modality alignment strategies in representation space, modeling the inter-modal relationships as a unified graph before representation learning leads to more accurate modality information transition, since it captures the dependencies among diverse modalities by directly considering them as nodes and edges.

This research introduces **MEGA** (Multimodal Neural Network with Efficient Graph Alignment), a novel framework designed for knowledge extraction in social media posts by effectively bridging the gap between visual and textual relations. MEGA's core innovation lies in its sophisticated graph alignment method. This method leverages both structural similarity and semantic agreement between visual objects in an image and textual entities in a sentence to establish correspondences. This distinguishes MEGA from previous multimodal approaches that rely on simple concatenation of graph representations using graph convolutional networks. By identifying the most similar nodes across visual and textual graphs based on structural and semantic features, MEGA achieves superior alignment of visual and textual relations.



Sentence: *JFK* and *Obama* at *Harvard*.
Named Entities: *JFK (PER), Obama (PER), Harvard (ORG)*

Text-based Methods:
(Residence of, *JFK, Harvard*)
(Residence of, *Obama, Harvard*)
(Spouse, *JFK, Obama*)

Object-level Multimodal Method:
(Member of, *JFK, Harvard*)
(Member of, *Obama, Harvard*)
(Siblings, *JFK, Obama*)

Concept-driven Multimodal Method:
(Graduated at, *JFK, Harvard*)
(Graduated at, *Obama, Harvard*)
(Alumni, *JFK, Obama*)

Figure 1.3: An example of hierarchical multimodal knowledge extraction. Compared with text-based methods, multimodal methods can extract relations with the guidance of visual contents. Concept-driven method bridges the semantic gap between low-level object features and high-level concept features and achieves better results.

1.2.2 A Hierarchical Representation Learning Method

When using auxiliary modality information, the core limitation of existing multimodal pretraining models arises from the fact that they only focus on object-level or image-level features, ignoring that the data of each modality exhibit hierarchical structure across individual object elements. Simple concatenation of low-level object features and textual representations proves insufficient for modeling relations involving higher-level semantics. For instance, consider Figure 1, where an object-level multimodal method incorrectly predicts ”member of” and ”siblings” as relations based on extracted visual objects like ”cap,” ”book,” ”uniform,” and ”man.” This inaccuracy stems from the semantic disparity between low-level visual features and the high-level textual relations they aim to represent. Drawing inspiration from visual concept detection, this research posits that bridging this semantic gap necessitates the incorporation of high-level concepts extracted from images. In the example of Figure 1, integrating concepts like ”education,” ”university,” and ”graduation” enables the accurate prediction of the correct relations – ”graduated at” and ”alumni.”

To address this challenge, this work introduces **RECK** (REtrieval with Cross-modal Knowledge), a novel multimodal knowledge extraction model that leverages external concept knowledge graphs to bridge the semantic gap between vision and language. RECK exploits the inherent hierarchical structure of knowledge graphs, where knowledge paths connecting low-level semantic nodes often traverse through high-level semantic nodes. These knowledge paths serve as bridges, enriching the semantic representation and facilitating accurate knowledge extraction.

1.2.3 A Robust Data Augmentation and Estimation System

Multimodal language understanding has garnered significant interest due to its ability to enhance semantic understanding by leveraging cross-modal inference. Notable examples include methods for Multimodal Named Entity Recognition (MNER) and Multimodal Relation Extraction (MRE), both of which capitalize on collaborative reasoning based on aligning textual and visual content. However, a critical challenge arises from the prevalence of misalignment between images and text in commonly used datasets, such as TRC and Twitter100k, where misalignment rates can reach as high as 60%. This misalignment introduces noise that can mislead model training and degrade performance.

Building upon this issue, this work introduces **TMR** (Translation Motivated Multimodal Representation learning), a framework that generates divergence-aware cross-modal representations. TMR achieves this by incorporating two key components: Generative Back-translation, which generates synthetic data to address misalignment, and High-Resource Divergence Estimation, which quantifies and accounts for the degree of divergence between modalities. This approach provides a robust and effective means to mitigate the negative impact of misalignment, enhancing the reliability and performance of multimodal language understanding models.

1.2.4 An Iterative Refined Graph Reasoning Application

Existing inductive reasoning schemes, where agent opinions are gleaned from disparate concepts at the word level and consensus is sought through bottom-up summarization, often encounter two critical limitations: the trivialization of opinions and focus diversion. While effective in confined natural language processing tasks with limited conceptual scope, this inductive approach falters in multimodal scenarios. The information-rich nature of certain modalities, particularly images, increases the likelihood of introducing distracting concepts, amplifying semantic divergence within the context and escalating the potential for trivialization. Moreover, employing Chain-of-Thought (CoT) reasoning in such scenarios can further exacerbate focus diversion by amplifying the impact of potentially biased newly introduced concepts.

To address these challenges, This research proposes a novel deductive reasoning scheme called **BDoG** (Blueprint Debate on Graph). Inspired by real-world blueprint debates, which emphasize the evaluation and refinement of a proposal (the blueprint) to address specific issues, BDoG adopts a top-down reasoning approach. It begins by aggregating concepts from various modalities and incorporating their relationships into an initial graph, serving as a blueprint that delimits the scope of discussion and prevents the infiltration of irrelevant semantics. Crucially, BDoG conducts the debate by marking down conclusions directly on the graph, thereby preserving specific concepts and mitigating the risk of trivialization through the merging of concepts into generalized representations. This deductive approach ensures a more focused and coherent reasoning process, effectively mitigating the challenges of trivialization and focus diversion inherent in inductive multimodal reasoning schemes.

1.3 Outline of the Thesis

This thesis explores the frontiers of multimodal knowledge extraction and reasoning, addressing key challenges and proposing novel solutions within a unified framework. Figure 1 provides a visual overview of the research landscape and the contributions of this work. The thesis introduces a comprehensive graph learning framework for multimodal knowledge extraction and explores three crucial strategies to tackle the challenges of hierarchical learning, bias mitigation, and iterative refinement. The remainder of this thesis is structured as follows.

Chapter 2 reviews the foundations of traditional knowledge extraction techniques and delves into the advancements in multimodal learning methods relevant to knowledge extraction and reasoning. It provides a comprehensive overview of the state-of-the-art and sets the stage for the subsequent contributions of this thesis.

Chapter 3 introduces MEGA (Multimodal Neural Network with Efficient Graph Alignment), a unified graph learning framework designed for efficient and effective knowledge extraction from multimodal data. The core of MEGA lies in its novel graph alignment strategy, which leverages both semantic and structural similarities to establish robust correspondences between visual and textual modalities.

Chapter 4 tackles the challenge of semantic inconsistency between vision and language, advocating for a hierarchical multimodal learning approach. It introduces RECK (RElation extraction with Cross-modal Knowledge), a method that utilizes external concept knowledge graphs to bridge the semantic gap by establishing connections between visual and textual concepts through semantically rich knowledge paths. Further enhancing RECK, a graph attention mechanism is incorporated to model multi-grained multimodal information within relevant subgraphs.

Chapter 5 addresses the problem of bias and distractions arising from data misalignment in multimodal knowledge extraction. It presents TMR (Translation Motivated Multimodal Representation learning), a robust and reliable technique inspired by the principles of back-translation in machine translation. TMR leverages generative models, such as stable diffusion, to generate additional visual content that complements the missing semantics in misaligned data. Additionally, a pretrained vision-language model is employed to estimate the divergence caused by misalignment, facilitating precise modality fusion.

Chapter 6 extends the utility of extracted multimodal knowledge graphs to the realm of multimodal reasoning. It proposes a novel approach that incorporates multi-agent debate to iteratively refine the initial static and coarse knowledge graph (the blueprint). The collaborative and competitive interactions between agents foster a deeper understanding of the knowledge, leading to more refined and insightful reasoning outcomes.

The final chapter concludes the thesis, summarizing the key contributions and highlighting potential future research directions in the field of multimodal knowledge extraction and reasoning. It underscores the significance of the proposed framework and strategies in advancing the understanding and utilization of multimodal data for knowledge acquisition and intelligent reasoning.

Chapter 2

Literature Review

In this chapter, we provide a comprehensive review of the literature related to multimodal knowledge extraction and reasoning tasks. We will further discuss on the advanced multimodal techniques that we utilized in this thesis. This taxonomy will help elucidate the relationships between various contributions and highlight the advancements in these fields.

2.1 Multimodal Knowledge Extraction and Reasoning

2.1.1 Traditional Knowledge Extraction Methods

Named Entity Recognition

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) that involves identifying and classifying named entities in text, such as people, locations, organizations, and products. NER plays a crucial role in various applications, such as information retrieval [109], question answering [72], and sentiment analysis. It has gained significant attention from researchers due to its impact on downstream NLP tasks, such as relation extraction and entity linking.

Neural models have been proposed and achieved state-of-the-art performance in various domains and datasets. For instance, the End-to-End Neural Entity Recognition model proposed by Ma and Hovy [76] achieved state-of-the-art performance on the CoNLL 2003 dataset. Similarly, the neural architecture proposed by Devlin et al. [22], known as BERT (Bidirectional Encoder Representations from Transformers), has achieved state-of-the-art results on various NLP tasks, including NER.

However, recognizing named entities in social media is challenging due to the short and noisy nature of social media posts. For instance, tweets are limited to 280 characters, and users often use slang, misspellings, and abbreviations. These factors lead to a significant deterioration in NER performance on social media.

To address this challenge, researchers have proposed various models that incorporate tweet-specific features such as at-mentions, hashtags, URLs, and emotions obtained using a new labeling scheme. For instance, Gimple et al. [26] proposed a model that incorporates tweet-specific features to improve NER performance on social media. Ritter et al. [86] proposed a T-NER system that uses LabeledLDA to exploit Freebase dictionaries as a source of distant supervision. However, their method only identifies whether a span is an entity or not.

Recent studies have reported performance gains by leveraging external sources of information, such as lexical information and several preprocessing steps. For instance, Baldwin et al. [8] proposed a shared task that aims to improve NER performance on social media by leveraging external sources of information. Similarly, Aguilar et al. [5] proposed a multi-channel neural architecture that uses multiple sources of information to improve NER performance.

Moreover, people frequently share their daily lives in social media using text and image posts. Visual content can assist in recognizing named entities. Moon et al. [80]

and Zhang et al. [123] proposed leveraging visual information to extract entities, and the visual content and textual representations are related by an attention mechanism. However, these models represent an image with a single vector trained with only one semantic label, which limits their ability to recognize multiple entities with different types.

Relation Extraction

Relation extraction is a critical task in Natural Language Processing (NLP), which aims to identify and classify the semantic relationships between entities in text. This task is essential in constructing a knowledge graph, which represents entities and their relationships in a structured format. In recent years, there has been significant progress in relation extraction due to the advancements in neural network-based methods.

Early approaches to relation extraction were based on statistical methods, such as kernel-based and distant supervision methods. However, these methods have limitations in handling complex relations and suffer from low recall and precision rates. Sequence-based methods have been proposed to address these limitations. Convolutional neural networks, recurrent neural networks, and transformers have been utilized to improve relation extraction performance. For instance, Wang et al. [104] proposed a CNN-based model for relation extraction, which achieved state-of-the-art performance on the SemEval 2010 Task 8 dataset [40]. Similarly, Zhang et al. [125] proposed a position-aware attention-based RNN model that achieved competitive performance on the same dataset. BERT, a transformer-based language model, has recently been used for pretraining and fine-tuning for relation extraction tasks, achieving significant improvements in performance.

Dependency-based models have also been proposed to incorporate structural information into predicting relations. These models use dependency parsing to capture syntactic and semantic relationships between entities. Compared to sequence-based methods, dependency-based models are better at capturing information from long distances, such as cross-sentence relations. Various studies have proposed dependency-based models [30], such as a framework for cross-sentence n-array relation extraction based on graph LSTMs, a graph recurrent neural network, and a path-centric pruning strategy with graph convolutional networks. Guo et al. [31] further improved the method by incorporating attentive graph weights.

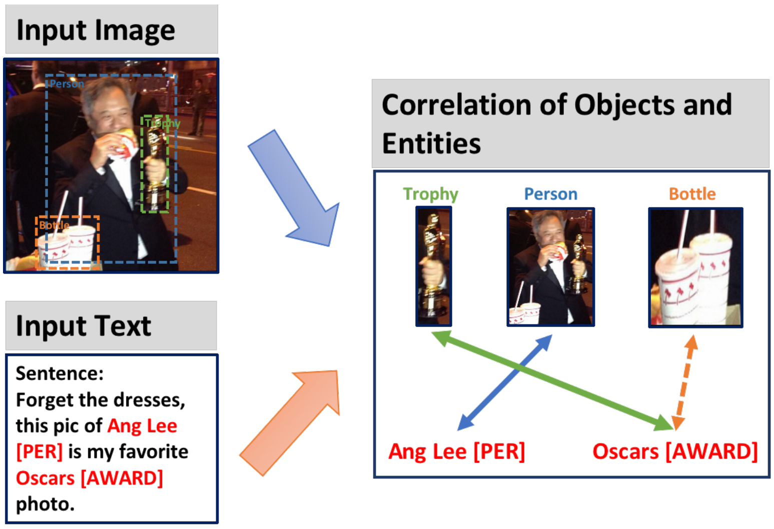
Despite the success of using dependency or external information, most existing methods suffer from performance decline when handling social media texts with sparse context. Social media posts, such as tweets, are typically short and contain informal language, slang, and emojis, which pose significant challenges for relation extraction. There is a need for more research in relation extraction on social media. Liu et al. [71] proposed a research direction for relation extraction on social media, and Brown et al. [10] discussed the challenges and errors in relation extraction on social media.

In conclusion, relation extraction is a crucial task in NLP that has gained significant attention from researchers in recent years. Neural network-based methods, such as CNNs, RNNs, transformers, and graph-based models, have shown promising results in improving relation extraction performance. However, there is a need for more research in handling social media texts with sparse context.

2.1.2 Multimodal Knowledge Extraction Methods

Multimodal Named Entity Recognition

Social media platforms such as Twitter and Instagram contain a vast amount of user-generated content, encompassing a variety of topics, including news, entertainment, and personal experiences. This data can provide valuable insights into human behavior, such as sentiment analysis, opinion mining, and trend analysis. Named Entity



Sentence:
Forget the dresses, this pic of Ang Lee [PER] is my favorite Oscars [AWARD] photo.

Figure 2.1: An example of the Twitter dataset. The visual object with label "person" will lead to the detection of "Ang Lee" as PER category, and objects with "trophy" will lead to the extraction of "Oscars" as the name of an award (MISC). The object "bottle" is irrelevant to entities in this post.

Recognition (NER) is a critical step in mining social media data as it allows for the identification and classification of named entities, such as people, organizations, locations, and products.

However, traditional neural-based NER models have limitations when it comes to social media data. Social media texts are usually short and informal, lack context, and are full of ambiguous expressions. For example, the sentence "I'm dying to try that new restaurant" could mean that the person is excited to try the restaurant or that they are literally dying and want to try the restaurant before they pass away. Therefore, to accurately identify named entities in social media data, additional methods are needed.

One approach to address this issue is to identify entities with external knowledge bases. For example, Ritter et al. [87] proposed a distantly supervised method that leverages a large amount of unlabeled data and large dictionaries to identify named entities. Li et al. [56] introduced an iterative method to split tweets into meaningful segments and evaluate the method on the NER task. However, these text-based methods rely solely on text data and cannot effectively identify named entities and their types when lacking textual context.

Therefore, current developments in deep learning and representation learning have led to the proposal of neural network-based multimodal NER methods. These methods utilize both image and text information for predicting named entities in social media. For example, Zhang et al. [123] proposed an adaptive multimodal method that combines the representations of visual objects and text to predict named entities. Moon et al. [80] proposed a multimodal neural network that learns to align visual and textual features for named entity recognition. Lin et al. [67] proposed a multi-task learning framework that jointly models named entity recognition and image classification for social media data.

However, these methods have limitations. The first limitation is that they ignore the mapping relations between visual objects and named entities. For example, in the sentence ”Ang Lee wins Oscars”, the visual object with the label ”person” is related to the named entity ”Ang Lee”, and the object ”trophy” is related to the named entity ”Oscars”. Previous multimodal NER methods representing the image with only one vector trained on one semantic label will mislead their models to extract different types of entities into the same type, resulting in inaccurate predictions.

Therefore, it is essential to utilize object-level features to distinguish entities with different types and extract entities accurately. For example, Lu et al. [72] proposed a visual-semantic embedding method that jointly learns image and text representations to capture the mapping relations between visual objects and named entities.

Another limitation is that previous works ignore the distribution disparity of image and text features. The distributions of image and text features are different, making it challenging to align named entities with image regions accurately. Therefore, a

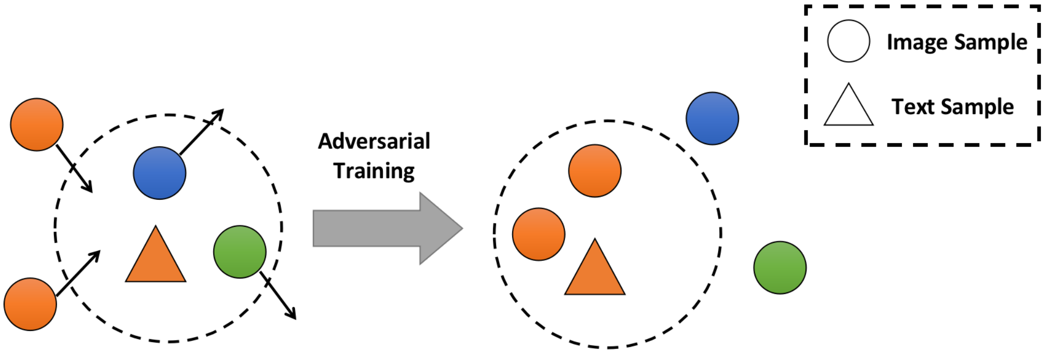


Figure 2.2: General idea to achieve an improved, modality-invariant subspace embedding with adversarial training. Shapes of the same color are semantically similar.

more effective method should be derived to bridge the distribution gaps for robust multimodal representations in the social media NER task. For example, Huang et al. [47] proposed a multi-attention network that learns to align image and text features at different granularities for named entity recognition.

In summary, social media data provides valuable information for understanding human behavior, and named entity recognition is a critical step in mining this data. Traditional neural-based NER models have limitations when it comes to social media data, and current developments in deep learning and representation learning have led to the proposal of multimodal NER methods that utilize both image and text information. However, these methods have limitations that need to be addressed to accurately identify named entities in social media data.

Multimodal Relation Extraction

Relation Extraction (RE) is a natural language processing task that involves identifying and extracting relationships between named entities in a sentence. Named entities can be anything from people, organizations, and locations to products, events, and concepts. The task of RE plays a critical role in various applications, such as question-answering systems, information retrieval, and knowledge graph construction.

Traditional RE methods, such as kernel-based or embedding methods, rely on human-annotated data, which is time-consuming and challenging to generalize well. Therefore, researchers have proposed neural network-based methods that achieve great success in different feature extractors. These methods use a combination of convolutional and recurrent neural networks to learn contextual embeddings of entities and relations.

However, most of these methods focus on the newswire domain, where sentences are formal and complete. In contrast, social media posts are often short and lack context, making it challenging to identify relations accurately. Therefore, researchers have proposed distant supervision, which leverages the alignment of knowledge bases and texts in sentences to automatically annotate relations. Distant supervision is a semi-supervised learning approach that uses existing knowledge bases to label relation instances in text corpora. However, distant supervision suffers from the problem of wrong labeling, which is even worse when contexts are missing.

To address this issue, researchers have proposed multimodal methods that combine visual information with text to supplement the missing semantic information. Visual contents, such as images, can provide additional context to improve the performance of identifying relations. For example, in a sentence like ”Steve Jobs founded Apple,” the image of an apple can provide additional context to distinguish the named entity ”Apple” from the fruit.

Researchers have proposed several multimodal neural relation extraction models that use both visual and textual features to identify relations accurately. For example, Zhang et al. [124] proposed a multimodal network that uses both text and image features to extract relations. The network first extracts visual features from images and textual features from text, then combines them to predict the relation between named entities.

However, the alignment of vision and language is a significant challenge in multimodal relation extraction. The current methods rely on the simple concatenation of representation vectors, which cannot effectively model the relationship of higher level semantics. Therefore, researchers have envisaged a range of well-designed methods and resources for such a challenge that would boost the development of multimodal alignment towards a higher semantic level. For example, Li et al. [60] proposed a method that uses a visual-semantic embedding to capture the mapping relations between visual objects and named entities.

In summary, multimodal relation extraction is an emerging field that combines visual and textual features to identify relations accurately. Traditional RE methods rely on human-annotated data and are time-consuming and challenging to generalize. Distant supervision is a semi-supervised learning approach that leverages the alignment of knowledge bases and texts in sentences to automatically annotate relations. However, distant supervision suffers from the problem of wrong labeling, which is even worse when contexts are missing. Therefore, researchers have proposed multimodal methods that combine visual and textual features to supplement the missing semantic information. The alignment of vision and language is a significant challenge in multimodal relation extraction, and researchers have proposed various methods to address this challenge.

2.1.3 Multimodal Reasoning

Multimodal reasoning is a crucial component in the development of advanced artificial intelligence (AI) systems that aim to replicate human-like intelligence [73]. This type of reasoning enables AI systems to process and analyze information from various sources and forms, such as text, images, audio, and video, in a integrated and coordinated manner [81, 11]. The latest advancements in multimodal large language models, such as BLIP2 [58], KOSMOS [48] and LLaVA [68] have demonstrated significant progress in complex reasoning, as these models [126] now have the capability to generate step-by-step rationales prior to producing the final answer, following a chain-of-thought (CoT) manner. Zheng et al. [132] propose a duty-distinct prompting method wherein questions are decomposed into sub-questions to enable deep-layer reasoning. SCITUNE [43] and T-SciQ [103] aim to teach large language models to answer science questions via the generation of mixed rationales derived from both large pretrained models and human annotators. Chameleon [74] accomplishes complex multimodal reasoning tasks by integrating various external tools (*e.g.*, large language models, off-the-shelf vision models, and web search engines).

To mitigate the susceptible error in CoT reasoning, Shinn et al. [92] and Madaan et al. [77] employ model to reflect on task feedback signals that can induce better decision-making in subsequent trials. [131] exploit previously generated answer as hint to progressively guide towards correct answer. Although these methods effectively enhance the performance of LLM, they struggle to produce novel ideas once they have determined a response, as they rely solely on internal representations for generation [46]. Researchers are currently developing multi-agent collaborative systems to address above issues in pure-textual scenarios [115]. By designing these systems, large language models (LLMs) can work together to complete tasks or engage in productive debates by offering contrasting perspectives [64, 13, 24]. Zhang et al. [121] further reveal the collaboration mechanism from a social psychology view. This thesis represents an initial endeavor to expand upon this method to facilitate multimodal reasoning. By incorporating multiple perspectives from different multimodal language models, we can help address some of the limitations of individual models.

Prior research has investigated the integration of structured graphs, such as knowledge graphs (KGs), into large language models (LLMs) by embedding the knowledge into the underlying neural networks [65, 106]. Nevertheless, embedding KGs within LLMs may compromise the inherent explainability and adaptability associated with knowledge reasoning and updating [44]. To tackle these challenges, recent studies have put forth innovative solutions. Li et al. [62] propose an adaptive query generator, facilitating the creation of queries across various query languages (*e.g.*, SPARQL) to infer rationales. Wang et al. [102] devise a structured multi-round question-answering (QA) format, which extracts external knowledge and generates coherent reasoning traces grounded in precise answers. Sun et al. [95] introduce Think-on-Graph (ToG), a method that sequentially reasons over KGs to locate relevant triples, thereby supporting the LLM in predicting the final answer. In the context of multi-modal reasoning, CCoT [78] substitutes the rationale generation process with scene graph extraction to enhance the compositional capabilities of large multimodal models. KAM-CoT [79], on the other hand, incorporates external KGs during the two-stage training process, yielding state-of-the-art fine-tuning outcomes in multimodal reasoning. In contrast to existing methods that utilize static graphs, our proposed BDoG preserves the dynamics and precision of KGs through iterative updates of entities, attributes, and relationships, guided by a blueprint debate process.

2.2 Related Advanced Multimodal Learning Techniques

2.2.1 Multimodal Pre-Training

Vision-language pretraining models are large-scale pretrained models that are capable of learning universal cross-modal representations by combining the strengths of computer vision and natural language processing. Two of the most popular PTMs are BERT and ViT [61], which have demonstrated remarkable success in representative learning and have become a milestone in machine learning.

The success of PTMs in computer vision and natural language processing has led to the development of multimodal PTMs that aim to extend this representation learning to the multimodal domain. These models, such as VinVL [122], VIL-T [53], and DALL-E 2, can significantly improve the performance of downstream multimodal tasks, such as visual question answering and image captioning.

However, despite the remarkable success of these models, text-image misalignment has rarely been studied, even though it is critical in real-world applications. Text-image misalignment refers to the mismatch between the textual and visual information, which can occur due to various factors such as image quality, description quality, and context. As a result, bridging the gap between text and image is a crucial challenge that needs to be addressed for vision-language pretraining models to be more effective in real-world applications.

To address the challenge of text-image misalignment, recent studies have proposed various approaches for vision-language pretraining models. One approach is to use contrastive learning to align the text and image features in a shared embedding space. This can be achieved by using different types of contrastive losses, such as InfoNCE [83], SimCLR [15], and MoCo [37], which aim to maximize the similarity between the positive image-text pairs while minimizing the similarity between the negative pairs.

Another approach is to use cross-modal attention mechanisms that can selectively attend to relevant regions in the image and text. This can be achieved by using different types of attention mechanisms, such as self-attention, cross-attention, and multi-modal attention, which can learn to attend to relevant visual and textual information while filtering out irrelevant information.

Moreover, some studies have proposed to incorporate additional modalities, such as audio and video, to improve the alignment between text and image. This can be achieved by using different types of fusion techniques, such as early fusion, late fusion, and cross-modal fusion, which can combine the different modalities in a meaningful way to improve the performance of the vision-language pretraining models.

Overall, the development of these approaches has shown promising results in addressing the challenge of text-image misalignment, and it is expected that future research will continue to explore and improve upon these methods to help vision-language pretraining models become more effective in real-world applications.

One of the challenges in developing effective vision-language pretraining models is the lack of large-scale multimodal datasets that capture the complexity and diversity of real-world multimodal data. To address this, recent studies have proposed the creation of large-scale datasets, such as Conceptual Captions, VQA, and COCO, which provide a rich source of annotated data for training and evaluating vision-language pretraining models.

Another challenge is the computational cost of training large-scale vision-language pretraining models. To address this, recent studies have proposed various techniques, such as distillation, knowledge transfer, and model compression, which aim to reduce the computational cost of training and deploying these models without sacrificing their performance.

Moreover, vision-language pretraining models have shown promising results in a wide range of applications, such as image captioning, visual question answering, and image retrieval. For example, recent studies have shown that vision-language pretraining models can generate more semantically meaningful captions and achieve state-of-the-art performance in image retrieval and visual question answering tasks.

In summary, the development of effective vision-language pretraining models is a rapidly evolving field that has shown remarkable progress and potential in bridging the gap between text and image. As the field continues to advance, it is expected that vision-language pretraining models will become more effective and applicable in a wide range of real-world applications.

2.2.2 Multimodal Generative Models

Generative diffusion models are a powerful class of generative models that can be used for a variety of downstream applications, including image synthesis, video generation, and molecular generation. They work by modeling the diffusion process of noise through a sequence of steps, where the noise is progressively transformed into the desired output.

One of the key advantages of diffusion models is their ability to generate high-quality, photorealistic images that closely resemble real-world images. This has been demonstrated in recent studies, such as BigGAN [9] and StyleGAN [1], which use diffusion models to generate high-quality images that are visually indistinguishable from real images.

Moreover, diffusion models have also been used to generate other types of data such as videos and molecules. For example, the recent work of Ho et al. [42] proposed a diffusion-based method for video generation that can generate high-quality videos with realistic motion and appearance. In the field of chemistry, diffusion models have been used to generate molecular structures and predict chemical reactions.

Furthermore, diffusion models have also been used for data augmentation and denoising tasks. For example, the recent work of Sohl-Dickstein et al. [94] proposed a diffusion-based method for denoising images, which can recover high-quality images from noisy inputs. Additionally, diffusion models have been used for data augmentation tasks in computer vision, such as image inpainting and super-resolution.

Overall, the development of generative diffusion models has shown remarkable progress and potential in various domains, and it is expected that future research will continue to explore and improve upon these models to help advance a wide range of applications.

2.3 Conclusion

In conclusion, the literature on multimodal knowledge extraction and multimodal reasoning techniques reveals a rich and evolving landscape of research that addresses the complexities of processing and understanding information from multiple sources. The advancements in Named Entity Recognition and Relation Extraction have laid the groundwork for more sophisticated approaches that leverage multimodal data, particularly in the context of social media.

The integration of visual information into NER and RE tasks has shown promise in enhancing performance, but challenges remain in aligning and interpreting the relationships between different modalities. The emergence of multimodal pre-training, generative diffusion models, and multimodal reasoning further underscores the potential of combining diverse data types to improve AI systems' capabilities.

As the field continues to evolve, ongoing research will be essential in addressing the challenges of multimodal knowledge extraction, including the need for better alignment techniques, high-quality datasets, and interpretable models. By tackling these challenges, researchers can unlock the full potential of multimodal learning and knowledge extraction, paving the way for more intelligent and capable AI systems that can understand and interact with the world in a more human-like manner.

Chapter 3

A Unified Graph Multimodal Learning Framework for Knowledge Extraction

3.1 Introduction

The undertaking of knowledge extraction aims to ascertain the semantic connections between two entities in an utterance. Knowledge extraction plays a crucial role in many applications necessitating comprehension of relationships such as question answering [4] and knowledge base population [49]. The majority of extant knowledge extraction methods can be dichotomized into two categories: sequence-dependent models and dependency-dependent models. Compared with sequence-dependent models, dependency-dependent methods can capture long-distance semantic dependency and commonly achieve superior performance.

However, these methods primarily hinge on text and suffer a precipitous performance decline in social media posts lacking context. For instance, in an utterance

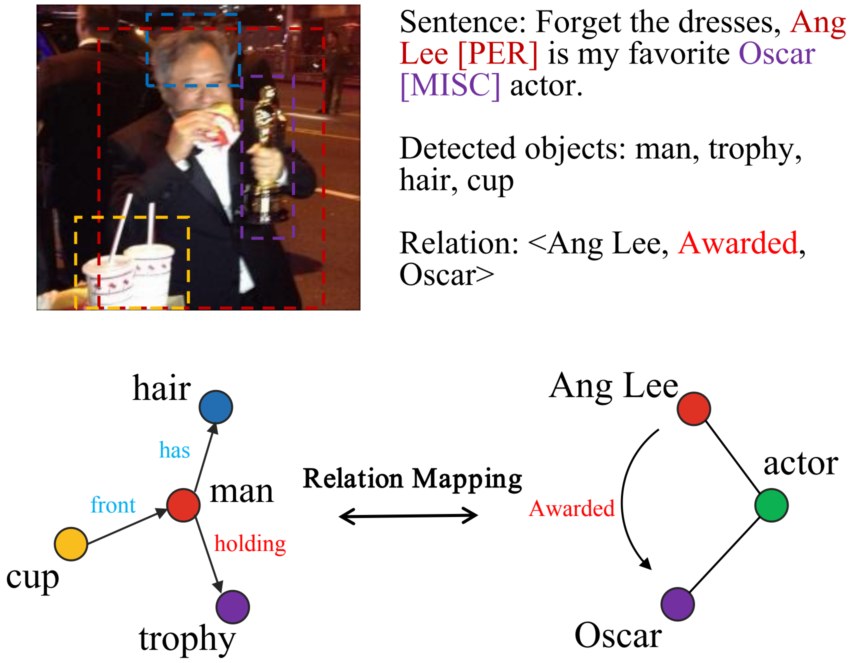

Sentence: Forget the dresses, Ang Lee [PER] is my favorite Oscar [MISC] actor.
Detected objects: man, trophy, hair, cup
Relation: <Ang Lee, Awarded, Oscar>
hair
has
man
front
cup
holding
trophy
Relation Mapping
Ang Lee
actor
Awarded
Oscar

Figure 3.1: An example of multimodal relation extraction in Twitter. The mappings from visual contents “man holding a trophy” to textual entities “Ang Lee” and “Oscar” will lead to the extraction of textual relation “awarded”.

”JFK and Obama at Harvard”, given two entities ”JFK” and ”Obama”, traditional methods can hardly detect the relation between them is ”Alumni” without other supplementary information. Consequently, most methods will incorrectly extract the relation ”couple” of the two entities since most cases in training data are labeled with such tags. We find that image-related information can supplement the missing context in relation extraction in social media texts. In the aforementioned case, we can effortlessly classify the relation into ”Alumni” with an image demonstrating that the two individuals don bachelor caps and the same school uniforms.

Exploiting visual contents to complement textual contexts has become a research hotspot in recent studies involving multimodal learning. Multimodal named entity recognition is one of the tasks necessitating comprehension of both vision and language. Zhang et al. [123] propose an adaptive co-attention network which utilizes image-level region features to assist extracting entities in tweets. Wu et al. [110] consider object-level features as fine-grained features and provide a novel attention method to align visual objects and textual entities. Dissimilar from the multimodal named entity recognition task, introducing visual information into knowledge extraction asks models not only to capture the correlations between visual objects and textual entities but also to focus on the mappings from visual relations between objects in an image to textual relations between entities in an utterance.

In this work, we study multimodal relation extraction (MRE), a specific task of knowledge extraction which classifies textual relations between two entities with the assistance of visual contents. Since there is no available dataset for training and evaluating MRE models, we present the MNRE dataset, a manually-labelled dataset for multimodal neural relation extraction. The corpus consists of texts and image posts crawled from Twitter. Four well-educated annotators were asked to tag both the entities and their relations. Owing to the noisy nature of social media texts and the limited characters of tweets, MNRE is a challenge dataset to test the multimodal representation, fusion and reasoning abilities of existing methods.

To learn the mapping from visual relations to textual relations, we propose a Unified Multimodal Graph Learning Framework - MEGA for relation extraction in social media posts. Following the success of dependency-dependent RE methods [10], we parse the utterances with a dependency tree tool. Considering scene graphs can represent images in a fine-grained manner and analyze relations with a graph structure, we apply a pretrained scene graph model to extract visual objects and their relations preliminarily. To capture the relation mapping from visual contents (”man holding a trophy”) to textual relations (”Ang Lee is awarded for Oscar”), we propose a graph alignment method that incorporates structural similarity and semantic agreement between visual objects in an image and textual entities in an utterance. Unlike previous multimodal methods simply concatenating the graph representations [123], our method can find the most similar nodes between two graphs with structural and semantic features, resulting in superior alignment of textual and visual relations. The corresponding visual relations can assist our model identify textual relations more precisely.

3.2 Problem Statement

Definition 1. ***(Multimodal Relation Extraction):*** *Given a social media post with text t and image v, and a pair of entities (e_1, e_2) in t, extract their relation $r \in R$ where R is a predefined set of relation types.*

3.3 The Proposed Framework: MEGA

In this section, we introduce the MEGA model for multimodal relation extraction, which is shown in Figure 3.2. In order to build the model, our work can be summarized as the following steps: (1) First, we extract the textual semantic representations with a pretrained BERT encoder. Besides, we generate the scene graphs from images which provide rich visual information including visual objects features and visual relations among the objects. To represent the semantics of images, we regard the object features in the extracted scene graph as the visual semantic features. (2) Secondly, to acquire the structural representations, we obtain the syntax dependency tree of the input texts which models the syntax structure of textual information. The visual object relations extracted by scene graph can be constructed as a structural graph representation. (3) Thirdly, to make good use of image information for multimodal relation extraction, we respectively align the structural and semantic information of multimodal features to capture the multi-perspective correlation between multimodal information. Then, we effectively merge the two aligned results. (4) Finally, we concatenate the textual representations which represent the two entities and the aligned visual representation as the fusion feature of text and image to predict the relations of entities.

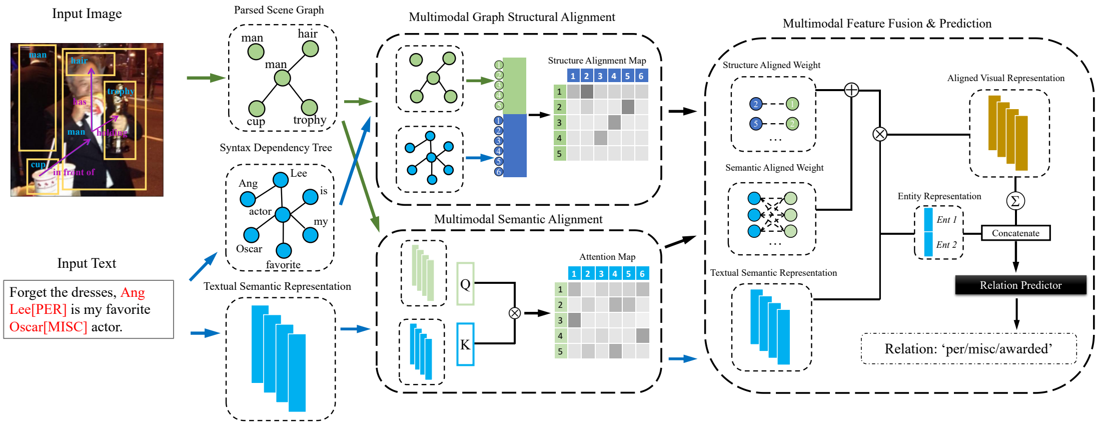


Figure 3.2: The Overall Framework of Our Proposed MEGA Model. Our Model Introduces Visual Information into Predicting Textual Relations. Besides, We leverages the Graph Structural Alignment and Semantic Alignment to Help Model Find the Mapping From Visual Relations to Textual Contents.

3.3.1 Semantic Feature Representation

Textual Semantic Representation

In the MNRE dataset, each piece of data contains a text message and an corresponding image from the social media posts, which is used as the input of our model. The input text message is first tokenized into a token sequence s_1. Then, to fit the BERT encoding procedure, we add the token '[CLS]' to the head of the sequence and the token '[SEP]' to the tail as well. In addition, following Soares et al.[93], we augment the s_1 with four reserved word pieces, $[E1_{start}]$, $[E1_{end}]$, $[E2_{start}]$ and $[E2_{end}]$ to mark the begin and end of each entity mentioned in the relation statement and modify s_1 to sequence $\tilde{s_1}$ as shown in Eq. (3.1),

$$\begin{aligned} \tilde{s_1} =& [w_1, ..., [E1_{start}], w_i, ..., w_{i+n_1-1}, [E1_{end}] \\ & , ..., [E2_{start}], w_j, ..., w_{j+n_2-1}, [E2_{end}], ..., w_l] \end{aligned} \tag{3.1}$$

where i and j denotes the start position of the first and second entity respectively. n_1 represents the length of the first entity while n_2 denotes the length of the second one. Besides, the token sequence are trimmed to a maximum length l. We pad the sample sequence which has less than l tokens to maximum length by [PAD] token.

Besides, we set a segment sequence to represent the segmentation of the valid tokens and [PAD] tokens. The segment sequence can be denoted as $s_2 = (1, 1, ..., 1, ..., 0, 0)$, where 1 represents the token which is not a padding one, 0 represents the [PAD] token. Therefore, the length of s_2 is l the same as $\tilde{s_1}$.

Following the success of Lample et al. [54], Ma and Hovy [75] , we represent each word in a input text message by combining character embedding into word embedding to obtain its textual features. We fine-tune the pre-trained BERT to get the embedding for each token in the sequence. The two sequences $\tilde{s_1}$, s_2 are fed into BERT to generate the embeddings. After that, each word is further transformed into a vector of d_x dimensions. And we can obtain the textual semantic representation by transforming the whole text message into a matrix $X \in \mathbb{R}^{l \times d_x}$, which is denoted in Eq. (3.2),

$$X = BERT(\tilde{s_1},\ s_2) \tag{3.2}$$

where $BERT$ denotes the BERT Encoder.

Visual Semantic Representation

Object-level visual features are considered as bottom-up manners in several multimodal tasks [6] to represents the image information. Therefore, we obtain the visual semantic feature by extracting the objects representation to represents the semantic of input image. In order to extract the objects from images, the input image is fed into the pre-trained scene graph generation model(with Faster R-CNN[85] as its backbone) to generate the scene graph of input image. An scene graph contains several nodes and edges connecting related nodes. The node contains the object features as its inner information, while the edges model the visual relation such as holding and wearing between different objects. In order to assist the entities relation extraction, we exploit the effective visual information while ignoring the irrelevant ones. Thus, we solely consider the top m salient objects with the higher object classification scores as the valid visual objects for further processing.

The input image is represented as a set of regional visual features in a bottom-up manner contained in the extracted scene graph. Each regional visual feature represents an object in the image with a vector y_i in dimension d_y . We set a confidence threshold to the probabilities of detected objects and obtain the top m objects for each image. Finally, an input image is transformed to a matrix Y. If the number of detected objects in an image is less than m, we would zero-pad Y to the maximum size m.

$$Y = [y_1, y_2, ..., y_m]_{m \times d_y} \tag{3.3}$$

3.3.2 Structural Feature Representation

In some previous works, the structure of the sentences (i.e., dependency trees) can provide important information which is benefit for the relation extraction models. Inspired by this, we generate two unidirectional graphs for the input text and image by using syntax dependency tree and scene graph generation model, which can provide the structural information to help multimodal relation extraction. It is notable that the visual object features plays the role as the node features in the scene graph.

Syntax Dependency Tree

Dependency tree is a structure used to express the dependency between words in a sentence. It has been shown in many previous work that the dependency trees can provide important information/features for the relation extraction. Each dependency

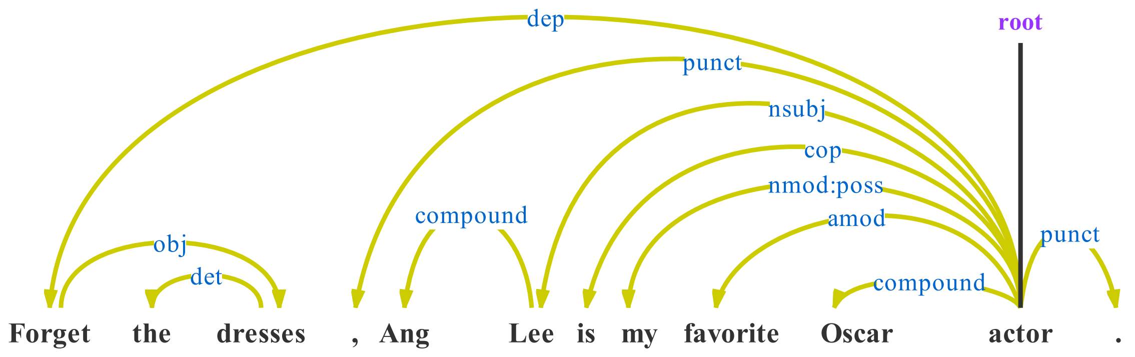


Figure 3.3: The Input Text Message is Performed by Syntactic Dependency Parsing. The Word *actor* is the Root Node of Dependency Relations while the Words in Blue (e.g., dep, obj) are Dependency Relations. The Direction of Arrow Indicates that There is a Relation Between the Two Words.

corresponding to two words from a sentence can be represented as a triple as Eq. (3.4):

$$R_{dependency} = (w_g, r_{type},\ w_d) \tag{3.4}$$

where w_g is the governor, w_d is the dependent and r_{type} shows how the dependent modifies the governor. We use ELMo [84], a common dependency tree extraction tool to obtain the dependency tree for the input text after which each word from the text is connected by its governor and obtains its related dependency triple. For example, the sentence *Forget the dresses, Ang Lee is my favorite Oscar actor.* is parsed to obtain the relations between words(e.g., amod, cop), as shown in Figure 3.3. The words in blue are the dependency relations. The ending of arrow indicates that this word is a dependent as well as a modifier. The word *root* in purple is used to indicate which word is the root node of dependency relations. Since each word is connected directly by another word in the text, the graph representation of the text is generated as G_1, which consists several relation pairs among the words.

$$V_1 = \{t_i | i \in [1, l_0]\} \tag{3.5}$$

$$E_1 = \{e_i = [t_i^*, t_i] | i \in [1, l_0]\} \tag{3.6}$$

$$G_1 = (V_1, E_1) \tag{3.7}$$

t_i represents the node corresponding to the ith token in the original text message which are not padded. t_i^* represents the governor of the ith token. l_0 represents the length of token sequence.

Scene Graph Generation

We obtain m objects and the visual relation between them from the input image by scene graph generation model. Since every relation between two objects is unidirectional, similar to the dependency tree, each object is also pointed by its governors from the image. Therefore, we can obtain the graph representation G_2 of the input image. G_2 consists several relation pairs of objects detected in the image and can be denoted as follows:

$$V_2 = \{o_j | j \in [1, m_0]\} \tag{3.8}$$

$$E_2 = \{e_{j,j_r} = [o_j, o_{j_r}^*] | j \in [1, m_0], j_r \in [0, r]\} \tag{3.9}$$

$$G_2 = (V_2, E_2) \tag{3.10}$$

where o_j represents the node corresponding to the jth object detected in the image. m_0 represents the number of detected objects. $o_{j_r}^*$ denotes the j_rth object which is related to jth object. $r \in [0, m_0 - 1]$ denotes the dynamic number of objects related to the jth object. After generating G_1 and G_2, we obtain the graph representation of the input text and image.

3.3.3 Multimodal Feature Alignment

To make full use of the obtained multimodal representation, we align the two graphs above from the structural perspective and use attention mechanism to align the textual and visual features from the semantic perspective.

Algorithm 1 Multimodal Graph Alignment

```
Input: Text graph $G_t$, Visual graph $G_v$
Output: Aligned graph $G_a$
Initialize alignment matrix $A \in \mathbb{R}^{|V_t| \times |V_v|}$
for each node pair $(v_t, v_v) \in V_t \times V_v$ do
    $A_{ij} \leftarrow \alpha \cdot sim_{struct}(v_t, v_v) + (1 - \alpha) \cdot sim_{sem}(v_t, v_v)$
end for
$G_a \leftarrow$ GraphFusion$(G_t, G_v, A)$ return $G_a$
```

Graph Structure Alignment

We exploit the node and edge information to extract the structure similarity of multimodal graph representation for structural alignment. First, as shown in Equation (7) and (10), we set $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ as two graphs mentioned above with node sets V_1 and V_2; edges sets E_1 and E_2 respectively. Let n be the number of nodes among two graphs, which means $n = |V_1| + |V_2|$. The steps of structure alignment can be summarized as follows: (1) obtain the node embeddings, conceptually by factorizing a similarity matrix of the node identities; (2) align nodes between two graphs by greedily matching the embeddings with an efficient data structure that allows for fast identification of the most similar embeddings from the other graph.

Following [39], we first set V_1 and V_2 into a union U shown as Eq. (3.11). In order to extract the node structural identity, we compute the counts of node degrees, including both in and out degrees of k-hop neighbors for each node u in U, which is shown as Eq. (3.11) and Eq. (3.12),

$$U = V_1 \cup V_2 \tag{3.11}$$

$$d_u^k = CountDegreeDistributions(R_u^k) \tag{3.12}$$

$$d_u = \sum_{k=1}^{K} \delta^{k-1} d_u^k \tag{3.13}$$

where $k \in [1, K]$, K is a graph diameter set by us and $\delta \in (0, 1]$ is a discount factor. And we compute the similarity between node a and node b in U as Eq. (3.14),

$$sim(a, b) = exp[-\gamma_s \cdot ||d_a - d_b||_2^2] \tag{3.14}$$

where γ_s is a scalar parameter controlling the effect of the structural identity. Then, we randomly select $p \ll n$ "*landmark*" nodes chosen across both graphs G_1 and G_2 and compute their similarities to all n nodes in these graphs using Eq. (3.15). This yields an $n \times p$ similarity matrix C, from which we can extract a $p \times p$ landmark-to-landmark submatrix W_p. Meanwhile, $W_p^\dagger$ is the pseudoinverse of W_p, a $p \times p$ matrix consisting of the pairwise similarities among the landmark nodes (it corresponds to a subset of p rows of C). It is a theorem that[39] given graphs $G_1(V1, E1)$ and $G_2(V2, E2)$ with $n \times n$ structural similarity matrix $S \approx PZ^T$, its node embedding matrix P can be approximated as

$$\tilde{P} = CU\Sigma^{1/2} \tag{3.15}$$

where $W_p^\dagger = U \sum V^T$ is the full rank singular value decomposition of the pseudoinverse of the small $p \times p$ landmark-to-landmark similarity matrix W_p. Now P and $\tilde{P}$ is the matrix with node embeddings as rows and its approximation. The p-dimensional node embeddings of the two input graphs G_1 and G_2 are then subsets of $\tilde{P}$: $\tilde{P}_1$ and $\tilde{P}_2$, respectively. We use Eq. (3.16) to obtain $\tilde{P}_1$ and $\tilde{P}_2$, which are the separate representations for nodes in G_1, G_2.

$$\tilde{P}_1, \tilde{P}_2 = D(N(\tilde{P})) \tag{3.16}$$

where D represents the dividing operation of $\tilde{P}$ by the number of $|V_1|$ and $|V_2|$ in order and N is used to normalize the magnitude of the embeddings and make them more comparable based on Euclidean distance. Finally, the last step is to efficiently align nodes using their representations, assuming that two nodes $i \in V_1$ and $j \in V_2$ may match if their embeddings in G_1, G_2 are similar. We find the alignments for each node by computing all pairs of similarities between node embeddings (i.e., the rows of $\tilde{P_1}$ and $\tilde{P_2}$) and choose the top-1 for each node. Here, we define the similarity a_{ij} between the p-dimensional embeddings of nodes i and j as follows:

$$a_{ij} = sim(\tilde{P_1}[i], \tilde{P_2}[j]) = e^{-||\tilde{P_1}[i]-\tilde{P_2}[j]||_2^2} \tag{3.17}$$

After the structural alignment of multimodal graphs, for each node in the text, its most similar node in structure from the image and their similarity score would be identified effectively. When finishing graph structure alignment, the two graphs are transformed into a feature matrix α,

$$\alpha = (\ a_{ij}\ )_{|V_1|\times|V_2|} \tag{3.18}$$

where a_{ij} represents the structural similarity between the ith word of the input text and the jth object of the input image. In our model, we only keep the most structurally similar object for each word while the elements corresponding to other objects except the most similar one are all represented by 0 in the matrix.

Semantic Features Alignment

In order to align the semantic of textual and visual information, we implement the guided-attention mechanism to capture the correlation between multimodal semantic features. The input of scale dot-product attention consists of queries and keys of dimension d_{key}, and keys of dimension d_{value}. For simplicity, we set d_{key} and d_{value} to the same number d_a. We calculate the dot products of the query with all keys, divide each by $\sqrt{d_a}$ and apply a softmax function to obtain the attention weights on the values. Given a query $q \in \mathbb{R}^{1\times d_a}$, n key-value pairs (packed into a key matrix $K \in \mathbb{R}^{n\times d_a}$ and a value matrix $V \in \mathbb{R}^{n\times d_a}$), the semantic aligned feature $y_a \in \mathbb{R}^{1\times d_a}$ is obtained by weighted summation over all values V with respect to the attention learned from q and K:

$$y_s \ = \ A(q, K, V) \ = \ softmax(\frac{qK^T}{\sqrt{d_a}})V \tag{3.19}$$

In practice, to obtain the semantic aligned features of all visual objects $Y_s \in \mathbb{R}^{m \times d_a}$, we compute the attention function on a set of m queries $Q = [q_1, q_2, ..., q_m] \in \mathbb{R}^{m \times d_a}$ seamlessly by replacing q with Q, which represents the visual semantic information guided by the textual features.

After obtaining the multimodal features representation, the input text is transformed into the matrix $X \in \mathbb{R}^{l \times d_x}$ and the input image is transformed into the matrix $Y \in \mathbb{R}^{m \times d_y}$. We employ three learnable matrix $W_k \in \mathbb{R}^{l \times d_a}$, $W_q \in \mathbb{R}^{m \times d_a}$ and $W_v \in \mathbb{R}^{l \times d_a}$ to generate the feature from X and Y for attention mechanism. In detail, the calculation process is shown from Eq. (3.20) to Eq. (3.22),

$$K = W_k X + b_k \tag{3.20}$$

$$Q = W_q Y + b_q \tag{3.21}$$

$$V = W_v X + b_v \tag{3.22}$$

where b_k, b_q, b_v are the learnable biases. As a result, we implement the semantic alignment by obtaining the semantic aligned weight β by calculation of Q and K as Eq. (3.23).

$$\beta = softmax(\frac{QK^T}{\sqrt{d}}) \tag{3.23}$$

Alignment Fusion

To fully use the structural and semantic alignment information, we integrate the aligned information by Eq. (3.24) to obtain the aligned visual features.

$$Y^* = (\alpha^T + \beta)V = \alpha^T V + Y_s \tag{3.24}$$

As we merge the structural and semantic alignment results, the final aligned visual features representation guided by the text is obtained as matrix $Y^* \in \mathbb{R}^{m \times d_a}$.

3.3.4 Relation Classification

To fully exploit the aligned visual information of all objects, we integrate the aligned object features to a vector representation, shown as Eq. (3.25),

$$\hat{y} = \sum_{i=1}^{m} y_i^* \tag{3.25}$$

where $y_i^* \in \mathbb{R}^{1\times d_a}$ represents the ith object feature in matrix Y^*.

Since we need to extract the relation between two entities from the text, we concatenate the representation $v_{[E1_{start}]}$ and $v_{[E2_{start}]}$ of their start position marker in feature V as the textual representation $\hat{v} \in \mathbb{R}^{1\times 2d_a}$ for multimodal fusion, which is shown as Eq. (3.26),

$$\hat{v} \;=\; [v_{[E1_{start}]}\;,\;v_{[E2_{start}]}] \tag{3.26}$$

We combine the guided visual information and the textual information from the two entities to obtain the final representation for the text and image by concatenating $\hat{v}$ and $\hat{y}$ into z, which is shown as:

$$z = concat(\hat{v}, \hat{y}) \tag{3.27}$$

Finally, we input z into an MLP to complete the final task of relation classification and obtain the output as shown in Eq. (3.28),

$$output = softmax(MLP(z)) \tag{3.28}$$

where $output \in \mathbb{R}^{n_c}$ represents the classification probability of all n_c relation categories.

3.4 Experimental Settings

3.4.1 Dataset

To provide empirical results for the effectiveness of our model, we construct a multimodal neural relation extraction dataset (MNRE) from scratch. The original corpus is built on three sources: two available multimodal named entity recognition datasets - Twitter15[72] and Twitter17[123], and crawling data from Twitter [1]. The posts were selected and filtered by annotators with different topics, such as music, sports and social events. We employed 12 well-educated annotators to label the relations between entity pairs and filter out the wrong samples tagged by automatic NER tools. The dataset contains 15,484 samples and 9,201 images with 23 relation categories. We split the dataset into training, development and testing set with 12247, 1624 and 1614 samples, respectively. The statistics of MNRE compared with a widely-used relation extraction dataset SemEval-2010 Task 8 [40] are listed in Table 3.1.

Table 3.1: The Statistics of MNRE Dataset Compared to SemEval-2010 Task 8 Dataset. # indicates Numbers.

Statistics	SemEval-2010	MNRE
# Word	205k	258k
# Sentence	10,717	9,201
# instance	8,853	15,485
# Entity	21,434	30,970
# Relation	9	23
# Image	-	9,201

We also show the distribution of relation categories in our MNRE dataset in Figure 3.4. We start tagging relation types depending on the entity types. For example, the

[1] https://archive.org/details/twitterstream

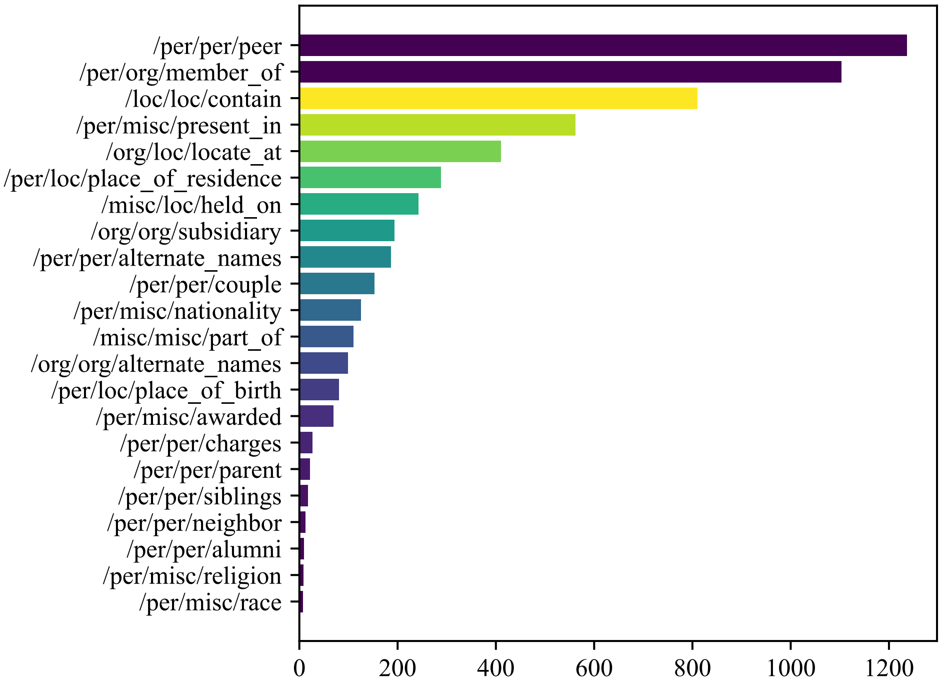


Figure 3.4: The Distribution of Relation Categories in Our MNRE Dataset.

relations between one person and another person can be classified into "alumni", "couple" and "relative" et al. We choose this labeling method since we expect the entity types and visual objects can be aligned and help to understand texts better.

3.4.2 Baseline Methods

We compare our methods with several relation extraction baselines. To validate the effectiveness of incorporating visual information into text-based RE models, we also provide several variants of the proposed MEGA model.

Glove+CNN Glove+CNN [118] is a classic CNN-based model for relation extraction. We use a improved version of this model [82] which concatenates word embeddings with position embeddings.

PCNN PCNN [117] is a distantly supervised relation extraction model which leverages external knowledge graphs to automatically label sentences with same entities contained.

Matching the Blanks (MTB) MTB [93] is an RE-oriented pretraining model based on BERT. Our method is built on the MTB model which, in turn, is the text-based version of the proposed MEGA model without visual features and the graph alignment strategy. We fine-tune it on our MNRE dataset as a text-based baselines.

BERT+SG The pretrained language model Bert [22] has shown its strong generalization in multiple tasks. We simply concatenate the fine-tuning BERT representations with visual features to show the improvement of introducing visual information. The visual features are extracted by a pretrained scene graph (SG) tool [97].

BERT+SG+Att. A variant of our proposed MEGA model which considers only the semantic similarity between visual graph (scene graph) and textual contents. Here we adopt the attention mechanism to compute the semantic similarity.

MEGA MEGA is our proposed multimodal relation extraction model with efficient graph alignment which considers both structural similarity and semantic agreement between visual and textual graphs.

3.4.3 Parameter Settings

We implement our model on the open-source and extensible relation extraction toolkit OpenNRE[34]which is based on PyTorch framework. To acquire the textual semantic representation, we initialize the textual representation by pretrained BERT and set the dimension d_x at 768. Besides, the dimension d_y of visual objects features extracted from scene graph is 4096. The latent dimension d_a of semantic alignment is set at 1536. The maximum number of token sequence and objects are 128 and 10 respectively. Our model is trained with Adamw optimizer, where we set the base learning rate at 2e-5 (following settings from previous works [34]) and the batch size at 10. The dropout rate in experiment is 0.5.

3.5 Results and Discussion

3.5.1 Overall Results

We conduct the experiments on the MNRE dataset. Table 3.2 shows the overall results on the test set of MNRE. We report accuracy, precision, recall and F1 value as the evaluation metrics. Compared to the traditional sequence-based CNN method [82], the distantly supervised RE model PCNN [117] achieves better results in all metrics. Since the MNRE dataset is collected with short social media texts, most words in training or testing set are novel words. In such case, a distantly supervised model will perform better with external KGs. However, the distantly supervised method will suffer the wrong labeling problem and the performance is restricted. Benefiting from the better generalization of pretraining language model representations, the MTB model [93] outperforms PCNN with a higher recall (64.46%) and F1 value (57.81%).

Table 3.2: The Overall Performance of Our Models and Other State-of-the-art Methods (Acc.: Accuracy, Prec.: Precision). * Indicates the Difference Against the F1 of Our Baseline Variant (MTB) is Statistically Significant by One-Tailed Paired *t*-test with $p < 0.01$.

Model	Acc.	Prec.	Recall	F1
Glove+CNN [82]	70.32	57.81	46.25	51.39
PCNN [117]	72.67	62.85	49.69	55.49
MTB [93]	72.73	64.46	57.81	60.96
BERT+SG	74.09	62.95	62.65	62.80*
BERT+SG+Att.	74.59	60.97	66.56	63.64*
MEGA	76.15	**64.51**	**68.44**	**66.41***

The other part of Table 3.2 is the performance of our MEGA model and its variants. All the variants of our methods outperform pevious text-based methods, which demonstrate the effectiveness of introducing visual information to supplement the missing text semantics. We use a pretrained scene graph parser to extract the fine-grained visual objects and their relations. Compared to simply concatenation of visual and textual features, a added semantic similarity module with attention mechanism will contribute to an improved recall value. We propose a more efficient alignment method which considers both structural and semantic similarity. Our model can align the visual and textual relations precisely and find more possible textual relations. As a result, the final MEGA model improves the precision and recall value (from 62.65% to 68.44%) in a large margin.

3.5.2 Performance on Categories

We also report the category results of our MEGA model compared to MTB model [93] in Table 3.3. Our model gains the highest results on all the six main categories on the MNRE test set. These categories involve the relations of person-to-person, person-to-organization or person-to-misc. Our model achieves relatively higher improvement in relation "Peer" and "Present_in". The two relations cover abundant visual information like "wearing the same uniform" or "appearing in a dance show". Our model introduces the visual information and utilize the mapping from visual relations to textual contents to help model extract relations precisely. However, text-based methods perform poor in these categories due to a lack of text contexts. The lower performance on location-related relations is due to: (1) Limited visual cues for location relations in images, (2) Implicit nature of location relationships requiring external knowledge, (3) Small number of training samples for these categories.

Table 3.3: Our Results on Six Main Categories Compared to MTB [93] on the MNRE Test Set.

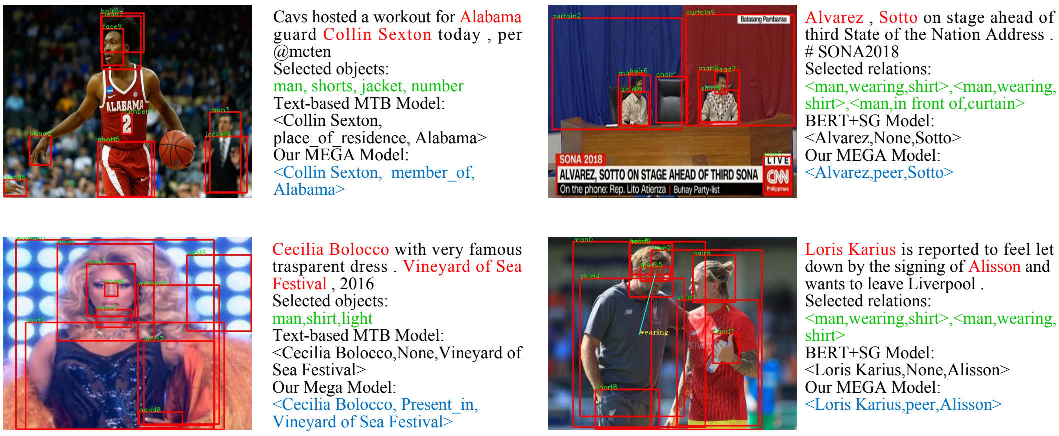
Category	Count	MEGA (Acc.)	MTB (Acc.)
Peer	156	**76.28**	63.46
Member_of	110	**70.90**	63.63
Contain	99	**91.91**	88.89
Present_in	74	**74.32**	51.35
Locate_at	46	**45.65**	41.30
Place_of_residence	29	**37.93**	31.03

Table 3.4: The Performance of Our MEGA Model on the MNRE Test Set Influenced by Different Number of Aligned Relations.

Aligned Relation Num.	Prec.	Recall	F1
Top-1	**64.51**	**68.44**	**66.41**
Top-5	64.13	64.53	64.33
Top-10	62.65	64.89	63.75

3.5.3 Parameter Sensitivity

Table 3.4 describes the results of our proposed MEGA model influenced by choosing different number of aligned relations. Top-1 indicates that for each word in a sentence, the most related visual object will be chosen. As mentioned before, we leverage the structural and semantic similarity to align the visual and textual features. However, there may be more than one visual relations related to textual contents. For example, we can ensure that two people are alumni with both “person wearing school uniform” and “person at the same school gate” visual contents. We find that in MNRE dataset, choosing the aligned relations with highest confidence will contribute to the best performance.



Cavs hosted a workout for Alabama guard Collin Sexton today , per @mcten
Selected objects:
man, shorts, jacket, number
Text-based MTB Model:
<Collin Sexton, place_of_residence, Alabama>
Our MEGA Model:
<Collin Sexton, member_of, Alabama>

Alvarez , Sotto on stage ahead of third State of the Nation Address . # SONA2018
Selected relations:
<man,wearing,shirt>,<man,wearing, shirt>,<man,in front of,curtain>
BERT+SG Model:
<Alvarez,None,Sotto>
Our MEGA Model:
<Alvarez,peer,Sotto>

Cecilia Bolocco with very famous trasparent dress . Vineyard of Sea Festival , 2016
Selected objects:
man,shirt,light
Text-based MTB Model:
<Cecilia Bolocco,None,Vineyard of Sea Festival>
Our Mega Model:
<Cecilia Bolocco, Present_in, Vineyard of Sea Festival>

Loris Karius is reported to feel let down by the signing of Alisson and wants to leave Liverpool .
Selected relations:
<man,wearing,shirt>,<man,wearing, shirt>
BERT+SG Model:
<Loris Karius,None,Alisson>
Our MEGA Model:
<Loris Karius,peer,Alisson>

Figure 3.5: The Results of Our Method (MEGA) Comparing to Text-based MTB [93] model and BERT+SG Model on the MNRE Test Set. Objects and Relations from Images are Detected in the Left Column, We Present the Relation Extraction Results with Related Objects and Visual Relations in the Right Column. The GroundTruth Labels are in Blue and the Detected Objects or Relations are in Green. Our Model Extracts Relations Precisely with Efficient Alignment between Images and Texts.

3.5.4 Case Study

Figure 3.5 shows the case study of comparing our MEGA model with the text-based MTB model [93] and BERT+SG model. With the help of efficient alignment between visual and textual relations, our model performs better in all cases. To evaluate the effectiveness of utilizing visual information, we compare our model with MTB model which only depends on textual information. On the left side of Figure 3.5, our model extract the relation "member_of" correctly with the guidance of visual objects "man, shorts, number". These objects indicate that the man is a player which is the member of a team. However, without the guidance of visual information, text-based method extracts the wrong relation "place_of_residence". Similarly, our model extracts the relation "present_in" with the guidance of visual objects "man, shirt and light" while the text-based method identifies it as "no relation".

On the right side of Figure 3.5, we compare our MEGA model with a variant of our model BERT+SG. BERT+SG model simply concatenates the visual and textual representations and ignores the mappings from visual relations to textual contents. For example, the BERT+SG model cannot classify the correct relation “peer”, however, our MEGA model finds the two people wearing the same uniform and extract the relation with alignment of visual and textual relations.

3.6 Conclusion

In this work, we present the multimodal relation extraction (MRE) task which leverages visual information to supplement the missing textual semantics in social media posts. To tackle this problem, we first provide a human-annotated dataset - MNRE which consists of 15000+ sentences with 23 relation categories. Then, we propose a multimodal relation extract neural network with efficient graph alignment (MEGA). MEGA uses graph-structured visual information to guide the extraction of textual relations with considering both structural and semantic graph similarity. The experimental results demonstrate that our model outperforms previous state-of-the-art methods in terms of precision, recall and F1 values.

Chapter 4

A Hierarchical Multimodal Representation Learning Method for Knowledge Extraction

4.1 Introduction

Given a pair of entities in a sentence, relation extraction (RE) task is to identify the relation between the two entities. Existing relation extraction datasets [125, 35, 40, 23] mostly concern on newswire domain where texts are formal and sentences are complete in semantics structures. However, these methods pose major limitations when sentences are too short to provide enough contexts, causing difficulties in figuring out the most possible relations, especially in social media. Figure 4.1 shows an example from Twitter where we may not predict the relations accurately between "Obama" and "JFK" or "Obama" and "Harvard" when the sentence gives no more external information. As a result, traditional RE approaches [118, 93] tend to determine the relationship between "JFK" and "Obama" as "None" based on category statistics. Vision and language (VL) learning draws significant attention in the past few years,

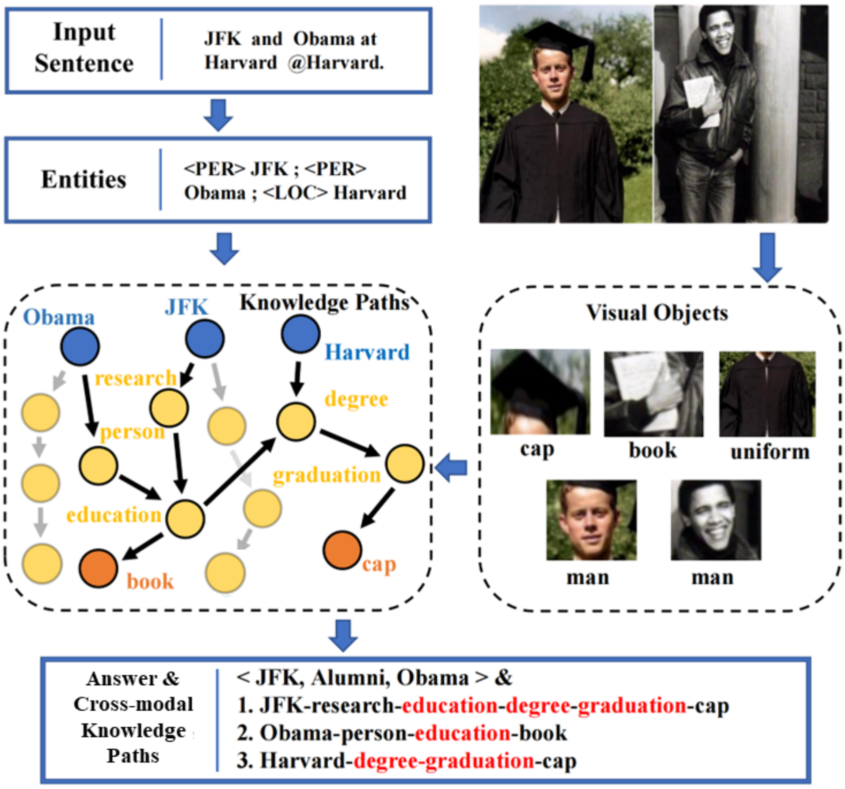

Input Sentence
JFK and Obama at Harvard @Harvard.
Entities
<PER> JFK ; <PER> Obama ; <LOC> Harvard
Knowledge Paths
Obama
JFK
Harvard
research
person
degree
graduation
education
book
cap
Visual Objects
cap
book
uniform
man
man
Answer & Cross-modal Knowledge Paths
< JFK, Alumni, Obama > &
1. JFK-research-education-degree-graduation-cap
2. Obama-person-education-book
3. Harvard-degree-graduation-cap

Figure 4.1: An example of multimodal relation extraction with cross-modal knowledge paths. We bridge semantic gaps between low-level visual objects and high-level textual relations. We also provide some direct evidences for relation extraction. As a result, our method can generate the correct relations with corresponding reasons.

with the proliferation of many VL benchmarks. Specifically, images in social media posts can provide complementary information to supplement the missing semantics of short texts.

Although some methods have been proposed to address the problem of multimodal RE [129, 128], there are several issues remaining challenging. When using auxiliary visual information, these methods only focus on visual objects, hence ignore that visual objects are quite specific and naturally in a lower semantic level than abstract textual relations. This inconsistency of semantic levels causes an incorrect alignment between vision and language, leading to final degrade of model performance. For example, in Figure 4.1, the ground-truth relation “Alumni” is quite abstract and has a semantic gap with specific objects “cap” and “book”. Therefore, these objects may not directly indicate that “JFK” and “Obama” are “Alumni”, resulting in the incorrect prediction.

Human beings can handle these challenges by developing proper imagination and reasoning around vision and language based on their prior knowledge. Motivated by this, many methods attempt to leverage external knowledge as supplementary information. However, existing knowledge-based methods model external knowledge as discrete points from single modality [109, 4], which inhibits the interactions between different modalities. As a result, visual information cannot be passed as a guidance to find reasonable knowledge for detecting textual relations. As shown in Figure 4.1, there are many knowledge paths among the two entities “Obama” and “JFK”. Without the guidance of visual contents, previous works tend to introduce much irrelevant information, and leading to inferior performance. Therefore, it is necessary to construct knowledge paths connecting two modalities with textual entities and visual contents.

In this work, we propose RECK - a multimodal relation extraction model with cross-modal knowledge representations. To address the semantic gap problem across vision and language, we propose to utilize the external concept knowledge graphs to build a bridge. We observe a phenomenon in the knowledge graph that there are some knowledge paths which might go through the high semantic level nodes when they connect two low semantic level ones. These knowledge paths can be used as the bridge to benefit the relation extraction. As shown in Figure 4.1, we construct several knowledge paths which include concepts “education”, “graduation” and “degree”. These concepts are closer to the correct relation “Alumni” in semantics. Considering that some points in the retrieved knowledge paths may be irrelevant or less important to our target relations, we apply a graph attention mechanism to filter out noisy visual information. Finally, we can fuse the knowledge representation with

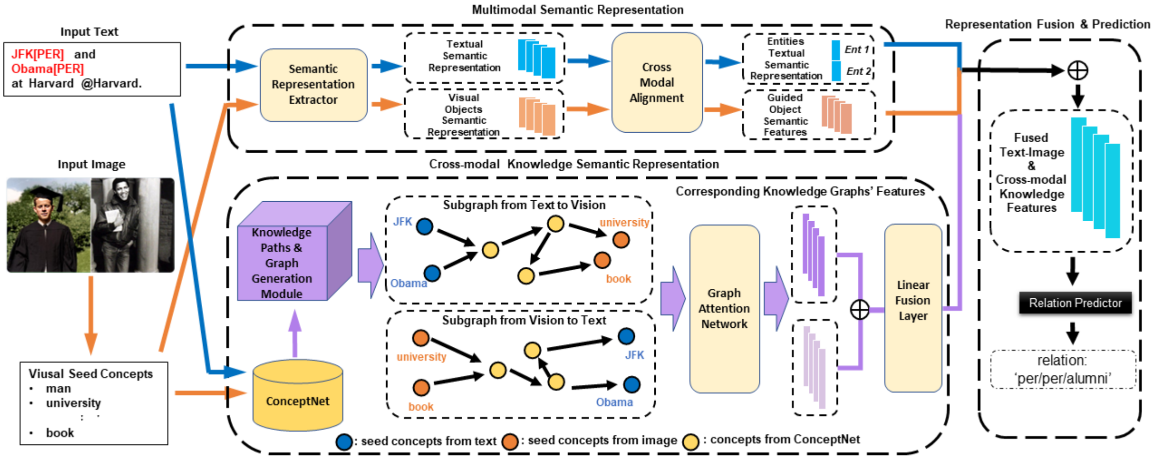

Input Text
JFK[PER] and Obama[PER] at Harvard @Harvard.
Input Image
Viusal Seed Concepts
• man
• university
⋮
• book
Multimodal Semantic Representation
Semantic Representation Extractor
Textual Semantic Representation
Visual Objects Semantic Representation
Cross Modal Alignment
Entities Textual Semantic Representation
Ent 1
Ent 2
Guided Object Semantic Features
Cross-modal Knowledge Semantic Representation
Knowledge Paths & Graph Generation Module
ConceptNet
Subgraph from Text to Vision
JFK
Obama
university
book
Subgraph from Vision to Text
university
book
JFK
Obama
Corresponding Knowledge Graphs' Features
Graph Attention Network
Linear Fusion Layer
●: seed concepts from text ●: seed concepts from image ●: concepts from ConceptNet
Representation Fusion & Prediction
Fused Text-Image & Cross-modal Knowledge Features
Relation Predictor
relation: 'per/per/alumni'

Figure 4.2: The overall framework of our proposed RECK. RECK consists of three main components, including multimodal semantics extraction, cross-modal knowledge semantics extraction and the module of fusion and prediction.

visual and textual features to identify textual relations.

4.2 Problem Statement

Definition 1. ***(Cross-modal Knowledge Path Learning):*** *Given text t, image v, and external knowledge graph G, construct knowledge paths $P = \{p_1, ..., p_k\}$ that bridge the semantic gap between visual objects $O = \{o_1, ..., o_m\}$ and textual entities $E = \{e_1, ..., e_n\}$.*

4.3 The Proposed Method: RECK

4.3.1 Overview

In this section, we introduce our proposed model RECK as shown in Figure 4.2. There are three components in RECK, including multimodal relation extraction and representation, cross-modal knowledge semantic extraction and representation, and representation fusion and prediction. The overall workflow of our model can be summarized into three steps: (i) we select knowledge paths from ConceptNet Database based on the text-image pairs, and further integrate the paths into knowledge-aware graphs, we then implement Graph Attention Network to extract the joint cross-modal knowledge representation for the integrated graphs, (ii) we employ the pre-trained BERT and ResNet to obtain textual and visual semantics representation for the text and image, and further adopt the cross attention mechanism to guide the model to learn the mixed multimodal features, (iii) we fuse all the feature representations to identify the relations among the entities.

4.3.2 Cross-modal Knowledge Representation Extraction

To bridge the semantic gaps, RECK extracts the cross-modal knowledge paths by following the flowchart of knowledge paths generation as shown in Figure 5.5.

Seed Concepts Extraction

To extract paths from ConceptNet, we need to generate query words for text and image first. These query words would be used as the start and end points of the paths. Following Chen et al. [14], we regard the query words as seed concepts. For the text input, since we want to bridge the semantic gaps for the textual relation between entities and the visual objects, we use the two entities as the textual seed concepts

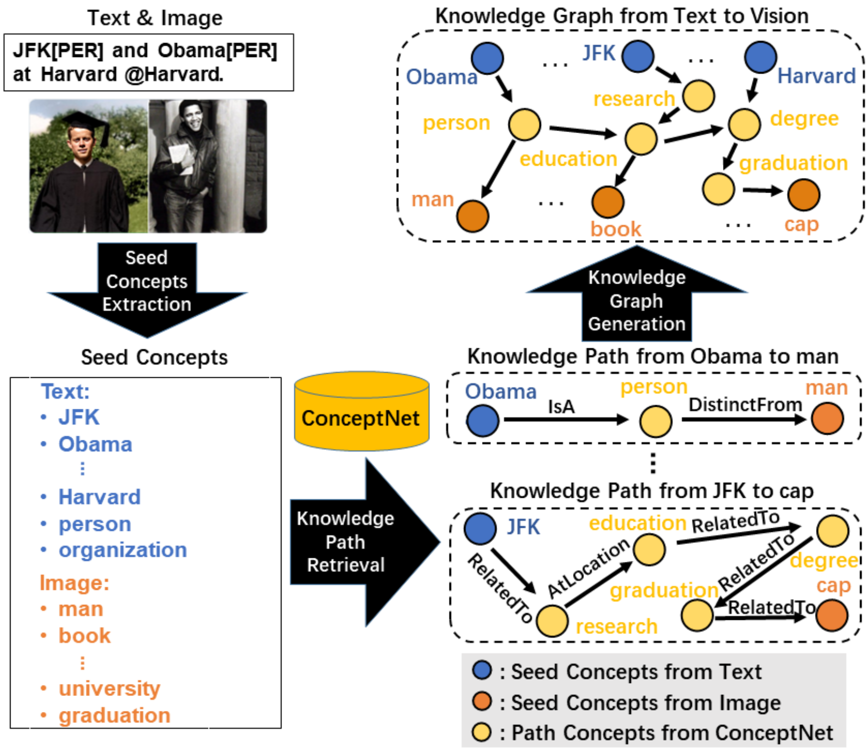


Figure 4.3: The procedure of the cross-modal knowledge path generation, taking the paths from text to vision as an example.

intuitively. In general, the case of words can affect the retrieval of ConceptNet. Thus, for uppercase words in texts, we keep their uppercase and lowercase states at the same time. As shown in Figure 5.5, “JFK” and “jfk” are both regarded as seed concepts. Besides, for the relation extraction task, the type of entity plays an important role in distinguishing the relationships among the entity-pairs. Therefore, we also regard the types as part of the seed concepts of text. For example, “person” is added to the seed concept candidates in Figure 5.5. As for other words of the sentence, not all of them are related to the target relation, so we do not regard them as the seed concepts. This is helpful to reduce noises brought from ConceptNet. Finally, the seed concept can be described as $t = [t_s, t_{sl}, t_{st}, t_o, t_{ol}, t_{ot}]$, where s, l, t, o represent subject (the first entity), the lowercase state, the type of entity and object (the second entity), respectively.

For the image input, we adopt a generic concept detection model *Clarifai*[1] to extract seed concepts. Compared to the frequently-used RCNN series model, Clarifai can not only detect objects of the image, but also provide words that describe the whole image. All these words will be regarded as seed concepts of the image. For example, “person” and “university” can be extracted from the same image by Clarifai. The words like “university” can help us better find a path in external knowledge to bridge the semantic gap and obtain better cross-modal knowledge representation, since more visual information is introduced. In particular, these words can hardly be detected from RCNN based models. In most cases, only the salient concepts are related to the MNRE task. Therefore, for each image, top k seed concepts with the highest score will be chosen, denoted as $c = [c_1, c_2, ..., c_k]$.

Knowledge Paths Retrieval

To bridge the gap among different semantic levels of different modalities, we select the external knowledge paths from ConceptNet Database based on the seed concepts as mentioned above.

We first expand the semantics bridging from text to vision. In details, we take every seed concept t from the input text as the starting point, and every seed concept c_i from the corresponding image as the end point. Then we search the top n shortest paths connecting the two concepts in ConceptNet as effective semantic bridging paths. Quite a part of the external concepts on the inference path can help fill the semantic level gap between low-level visual contents and high-level textual entity relations. For example, in Figure 4.1, when connecting “JFK” and “book”, word “education” would be introduced and it is closer to “alumni” at the semantic level. Besides, word “education” appears on the path from “JFK” to “book” and “Obama” to “book”, which provides indirect evidence for inferring that “JFK” and “Obama” are alumni.

[1] https://www.clarifai.com/

Reversely, we also take every seed concept c_i of the image as the starting point, and every seed concept t of the text as the end to form an opposite semantic bridging path. Because in ConceptNet, the connections of the nodes are directed edges, this indicates that building connectivity paths in both directions can bring more information and help fill in the semantic gaps from different perspectives. Finally, we denote the i^{th} knowledge path from text to vision as P_{t_i,v_i} and knowledge paths from vision to text as P_{v_i,t_i}, where t and v denote the text and vision modality, respectively.

Knowledge Representation Extraction

As described previously, we bridge the semantic gap by constructing knowledge paths. However, not all the concepts from the paths are relevant for the reason that ConceptNet inevitably introduces noise knowledge. Therefore, we need to find out the more positive and relevant concepts from the paths. In other words, it is necessary to re-evaluate the effect of each introduced concept on relation extraction.

In detail, we first integrate all possible knowledge paths extracted into a knowledge-aware graph, where each concept on a path denotes a node in the knowledge-aware graph. Then we adopt Graph Attention Network (GAT) [100] to re-weight the concepts relevance.

The concepts with higher weight scores indicate that they are more relevant to the relation extraction.

As we have obtained two kinds of knowledge paths previously, two kinds of knowledge-aware graphs can be generated, as depicted by Eq. (4.1):

$$KG_{(M_1,M_2)} = \bigcup_{i \in N_{M_1 M_2}} P_{M_{1i},M_{2i}} \tag{4.1}$$

The symbol KG_{M_1,M_2} denotes that the knowledge-aware graph consists of all the knowledge paths from modality M_1 to modality M_2. $N_{M_1 M_2}$ denotes the number of the corresponding knowledge paths. For instance, $KG_{v,t}$ indicates the knowledge-aware graph from vision modality to text modality and N_{vt} is the number of these knowledge paths. Such graphs contain abundant external knowledge with higher-level semantics, based on which the semantic gap between different modalities can be eliminated. To represent such a graph, we following Yang et al. [112] by choosing not to use the specific relation (*e.g.*, RelatedTo, Antonym) between any two concepts from the graph. Instead, for the i^{th} concept node of a graph, we first transform it into a knowledge graph embedding vector g_i supported by ConceptNet [69], which has fused the semantic information of edges. Thus, two graphs $KG_{t,v}$ and $KG_{v,t}$ can be transformed into two matrices: $G_{t,v} \in \mathbb{R}^{NC_{tv} \times d_l}$ and $G_{v,t} \in \mathbb{R}^{NC_{vt} \times d_l}$, where d_l represents the dimension of knowledge graph embedding, NC_{tv} and NC_{vt} represent the number of concepts in $KG_{t,v}$ and $KG_{v,t}$, respectively. As the following operations on a denoted graph are the same, we use G to denote any of the knowledge-aware graphs in subsequent discussion. We build an adjacency matrix $A = \{a_{ij} | a_{ij} \in \{0, 1\}\}$ for each graph, in which 1 represents that there is a directed edge between nodes i and j, while 0 means that the two nodes are not connected. We update G to G^* by concatenating G and A. Then, we employ a multi-head GAT to update the concept node features for each graph. Specifically, the attention weight between two concepts can be calculated by Eq. (4.2),

$$e_{ij} = SA(W_g g_i^*, W_g g_j^*) \tag{4.2}$$

where SA represents the self attention mechanism and W_g represents a learnable matrix. In order to make the coefficients easier to compare between different nodes, we normalize e_{ij} by Eq. (4.3),

$$\alpha_{ij} = softmax(e_{ij}) \tag{4.3}$$

After that, the normalized attention coefficient and its corresponding concept feature are linearly combined by the multi-head GAT to get the graph based representation $\hat{G}$, as denoted by the equation below:

$$\hat{g}_i = \overset{m}{\underset{k=1}{\|}} \sigma(\sum_{j=1}^{\hat{N}} \alpha_{ij}^k W^k g_j^*) \tag{4.4}$$

where σ is a activation function and W^k is a learnable matrix. $\hat{N}$ denotes the node number in $\hat{G}$ and m denotes the multi-head number of GAT.

Cross-modal Representation Generation

We integrate the feature representations of all nodes in each graph into a vector representation. As a result, the final representation vector $\tilde{g}$ of one knowledge-aware graph G is shown as Eq. (4.5),

$$\tilde{g} = \sum_{i=1}^{\tilde{N}} \hat{g}_i \tag{4.5}$$

where the vector $\tilde{g}$ can be regarded as the filter results of graph G.

Finally, to get the joint external reasoning information, we concatenate $\tilde{g}_{t,v}$ and $\tilde{g}_{v,t}$ and put the result into a fully connected layer, as shown in Eq. (4.6),

$$\tilde{g} = FC(concat(\tilde{g}_{t,v}, \tilde{g}_{v,t})) \tag{4.6}$$

where the FC represents the fully connected layer.

The result $\tilde{g}$ is the final cross-modal representation of each text-image pair.

4.3.3 Multimodal Semantic Extraction

Textual Semantic Extraction

For this process, any input text is first divided into a token sequence denoted as s_1. Following the criteria of BERT [21] encoding procedure and the success of Soares et al. [93], the token "[cls]" is added to the beginning of the sequence and the token "[sep]" is appended to the tail. Meanwhile, four more reserved word pieces, "$[E1_{start}]$", "$[E1_{end}]$", "$[E2_{start}]$" and "$[E2_{end}]$" are added into s_1 to record the begin and end of the first and second entity. Further more, to unify the length of all the sequences, we pad the sequence which has less than maximum length l tokens to l by adding "[pad]" tokens. Thus, sequence s_1 would be transformed into sequence $\tilde{s_1}$ as shown in Eq. (4.7),

$$\begin{aligned} \tilde{s_1} = & [w_1, ..., [E1_{start}], w_i, ..., w_{i+n_1-1}, [E1_{end}] \\ & , ..., [E2_{start}], w_j, ..., w_{j+n_2-1}, [E2_{end}], ..., w_l] \end{aligned} \tag{4.7}$$

where i and j represents the start position of the first entity and second entity, respectively, while n_1 and n_2 denotes the length of the two entities.

Since token '[pad]' is added to $\tilde{s_1}$, we also construct a segment sequence s_2 to distinguish between valid tokens and '[pad]' tokens, as shown in Eq. (4.8),

$$s_2 = (1, 1...1...0, 0) \tag{4.8}$$

where 1 represents that the token is not a padding one. Correspondingly, the length of s_2 is also l.

Inspired by Lample et al. [55], Ma and Hovy [76], we obtain each token's textual features by combining character embedding into word embedding. To achieve that, $\tilde{s_1}$, s_2 are fed into a fine-tuned pre-trained BERT to generate the embedding for each token. As a result, each token is further transformed into a vector x of d_x dimensions. Meanwhile, the textual semantic representation can be obtained by transforming the whole text into a matrix $X \in \mathbb{R}^{l \times d_x}$, shown as Eq. (4.9),

$$X \ = \ BERT(\tilde{s_1}, \ s_2) = [x_1, x_2, ..., x_l] \tag{4.9}$$

where $BERT$ denotes the fine-tuned pre-trained BERT encoder.

Visual Semantics Extraction

Visual objects are demonstrated as a fine-grained image representation and contribute to many multimodal tasks such as VQA [27], multimodal named entity recognition [130] and vision-language pretraining [57]. Following the great success of Chen et al. [17, 16], we apply a visual grounding tool [114] to extract three objects with highest detection scores from the image. Then, we extract the detection boxes corresponding to the objects and re-enlarge them to the size of the original image. Combined with the original one, we obtain four images for each instance. Similar to previous handling, all the images are denoted as c. To obtain the features representation, we feed c into the fine-tuned pre-trained ResNet50 [38]. By extracting the input features of the last layer of the model, we also generate a vector y of d_y dimensions for each image of the instance. As a result, the visual semantics representation can be obtained by transforming the images vector into a matrix $Y \in \mathbb{R}^{4 \times d_y}$, as shown in Eq. (4.10),

$$Y = ResNet50(c) \tag{4.10}$$

where $ResNet50$ denotes the fine-tuned pre-trained ResNet50 model.

Cross-modal Attention Mechanism

Inspired by the scale dot-product attention proposed in [99], we employ the cross-attention mechanism to find out the hidden correlation between textual and visual semantics representation. There are three kinds of input for guided attention: queries of dimension d_q, keys of dimension d_k and values of d_v. In particular, we set the three dimensions to be the same number d_g. We first calculate the dot products of the query with all keys. Then the product is divided by $\sqrt{d_g}$ and entered into a softmax function to obtain the guided-attention weights on the values. Besides, we pack n key-value pairs into two matrix $K \in \mathbb{R}^{n \times d_g}$ and $V \in \mathbb{R}^{n \times d_g}$. Given a query $q \in \mathbb{R}^{1 \times d_g}$, a guided feature $y_g \in \mathbb{R}^{1 \times d_g}$ for an image from c is obtained by weighted summation over all values V with respect to the attention learned from q and K, as shown in Eq. (4.11),

$$y_g = CA(q, K, V) = softmax(\frac{qK^T}{\sqrt{d_g}})V \tag{4.11}$$

where CA represents the cross attention mechanism.

Correspondingly, we compute Eq. (4.11) on a set of k queries $Q = [q_1, q_2, ..., q_k] \in \mathbb{R}^{k \times d_g}$ seamlessly by replacing q with matrix Q.

In the previous two sections, the text and image are respectively converted to the matrix $X \in \mathbb{R}^{l \times d_x}$ and matrix $Y \in \mathbb{R}^{4 \times d_y}$. To get the guided features of the image, we employ three learnable matrices $W_{kv} \in \mathbb{R}^{d_g \times d_x}$, $W_{qv} \in \mathbb{R}^{d_y \times d_g}$, $W_{vv} \in \mathbb{R}^{d_x \times d_g}$ and three corresponding bias $b_{kv}, b_{qv}, b_{vv} \in \mathbb{R}^{\eth}$ to help generate K_v, Q_v and V_v. After that, we can align the text-guide-image semantic information according to Eq. (4.12) to Eq. (4.15),

$$Q_v = Y{W_{qv}}^T + b_{qv} \tag{4.12}$$

$$K_v = X{W_{kv}}^T + b_{kv} \tag{4.13}$$

$$V_v = X{W_{vv}}^T + b_{vv} \tag{4.14}$$

$$Y^* = softmax(\frac{Q_v K_v^T}{\sqrt{d_g}})V_v \tag{4.15}$$

Correspondingly, in order to acquire the guided textual features, we also employ three more learnable matrices $W_{kt} \in \mathbb{R}^{d_g \times d_y}$, $W_{qt} \in \mathbb{R}^{d_g \times d_x}$ and $W_{vt} \in \mathbb{R}^{d_g \times d_y}$ and three corresponding bias $b_{kt}, b_{qt}, b_{vt} \in \mathbb{R}^{\eth}$ to help generate K_t, Q_t and V_t. And then, we can also align the image-guide-text semantic information by the following equations.

$$Q_t = X{W_{qt}}^T + b_{qt} \tag{4.16}$$

$$K_t = Y{W_{kt}}^T + b_{kt} \tag{4.17}$$

$$V_v = Y{W_{vt}}^T + b_{vt} \tag{4.18}$$

$$X^* = softmax(\frac{Q_t K_t^T}{\sqrt{d_g}})V_t \tag{4.19}$$

Inspired by BERT, we set up a twelve-layer cross attention mechanism to get the text-guide-image and image-guide-text feature representations.

4.3.4 Fusion and Prediction

Inspired by the success of Soares et al. [93], we concatenate the output semantic features of two entities from matrix X^* to represent the final semantics representation of the text, as shown in Eq. (4.20),

$$\tilde{x} \;=\; concat(x^*_{E1_{start}}, x^*_{E2_{start}}) \tag{4.20}$$

where $concat(*)$ represents the vector concatenation.

Besides, we integrate the aligned visual features to a vector $\tilde{y}$ to obtain the final visual semantic features, as shown in Eq. (4.21),

$$\tilde{y} = \sum_{i=1}^{k} y_i^* \tag{4.21}$$

In the end, we concatenate the textual features $\tilde{x}$, visual features $\tilde{y}$ and cross-modal knowledge-aware graph features $\tilde{g}$ as the final representation for the text-image pair, depicted by Eq. (4.22) bellow:

$$z = concat(\tilde{x}, \tilde{y}, \tilde{g}) \tag{4.22}$$

Finally, we set z into a multi-layer perception to realize the relationship classification of the entity pair. The output feature is shown as Eq. (4.23),

$$output = softmax(MLP(z)) \tag{4.23}$$

where MLP means multi-layer perception, $output \in \mathbb{R}^{n_r}$ represents the classification probability of all n_r relation categories.

4.4 Experimental Settings

4.4.1 Dataset

In this section we perform a set of experiments based on the MNRE [129] dataset, a manually-labeled dataset for multimodal relation extraction. The texts and image posts are crawled from three sources: two open-source multimodal named entity recognition datasets - Twitter15 [72] and Twitter17 [123] and the crawling posts from twitter[2]. To preserve the timeliness and diversity of data, they do not pick some certain topics but randomly choose samples with a large time span in crawling twitter data. As for the entities, they use a pretrained named entity recognition model[3] to extract entities.

The dataset contains 15,484 samples and 9,201 images with 23 relation categories. It is split into training, validating, and testing set with 12247, 1624 and 1614 samples, respectively. Relation types are tagged depending on the entity types. For example, the relations between one person and another person can be classified into “colleague”, “couple” and “relative”, etc.

4.4.2 Implementation Details

We implement our model on the open-source relation extraction toolkit OpenNRE [33] which is based on PyTorch framework. The text input is trimmed to the maximum of 128 words. We extract the textual features from pre-trained BERT with dimensions of 768, and the visual features from pre-trained ResNet50 with dimensions of 2048. As for the cross-attention mechanism, we set d_g as 768.

The concept features are initialized by 300-D ConceptNet KG embedding, includ-

[2] https://archive.org/details/twitterstream

[3] https://allennlp.org/elmo

ing the seed concepts from two modalities and external knowledge path concepts. About the GAT, the hidden and output dimensions are set at 512 and 768, respectively, and the number of head is set at 4. The dropout rate is 0.05 to avoid over-fitting in GAT. And we use a LeakyReLU optimizer with 0.01 alpha value for GAT.

Table 4.1: The Settings of HYPER-PARAMETERS.

Parameter	Value	Parameter	Value
max text length	128	BERT hidden dim d_x	768
ResNet50 hidden dim d_y	2048	CA hidden dim d_g	768
KG embedding dim d_l	300	GAT hidden dim	512
GAT output dim	768	GAT head number	4
GAT dropout rate	0.05	GAT alpha value	0.01
batch size	4	epoch number	8
optimizer	AdamW	learning rate	1e-5

For all the experiments, we set the batchsize at 16 and training epoches at 10. Besides, we use AdamW optimizer with the base learning rate of 1e-5. We select the seed concepts extracted from the image with top-5 degree of confidence and top-3 shortest paths between two seed concepts from text and vision. As same as other classification tasks, we adopt precision, recall, F1 score and accuracy as the evaluation metrics. In details, all the default setting of the model is depicted in Table 4.1.

4.4.3 Compared Baselines

We compare our model with several state-of-the-art (SOTA) neural relation extraction models based on single-modality and multi-modality.

- **GloVe+CNN** [82] is a classic CNN-based neural relation extraction model which combines the position embedding and word embedding to represent the textual sentence.

- **PCNN** [117] is a distant supervision model for relation extraction which captures the structural features between entity pairs more effectively.
- **MTB** [93] is a RE-oriented pre-trained model which implements the entity markers for textual relation extraction.
- **RECK(Single modal)** is the single-modal version of our model RECK, using only the text and the external knowledge introduced by entity seed concepts. In detail, we construct knowledge paths between two seed concepts extracted from two textual entities.
- **MTB+SG** [128] is a strong multimodal baseline which extracts the visual scene graph from to capture the correlation of visual objects and concatenates the multimodal features for prediction.
- **MEGA** [128] is a graph-based model which considers both structural similarity and semantics agreement between visual and textual graphs.
- **MKGformer** [16] is a multimodal model leveraging a hybrid transformer architecture with unified input-output for diverse multimodal tasks, including MRE.
- **HVPNeT** [17] is a model which works for MNER and MRE, regarding visual representation as pluggable visual prefix to guide the textual representation for error insensitive forecasting decision.

Besides, we also evaluate the performance of pre-trained multimodal model in the multimodal relation extraction task. **ViLT** [53] is a pre-trained multimodal model which cuts the image into patches and captures the modality interaction by Transformer-based structure. Besides, Chen et al. [16] adopt the pre-trained multimodal model VisualBert [61] on MRE task. We also introduce their results for comparison.

4.4.4 Ablation Experiments Setting

To fully evaluate the impact of each module of our model, we set up ablation experiment for each module module separately. In details, the ablation setting is shown as follows:

- **RECK(text only):** This model solely relies on textual input in RECK, indicating that only the semantic representation of text is utilized for relation classification, and there is no incorporation of image or external knowledge.
- **RECK(image only):** This model exclusively uses image information from the instances to perform relation classification and does not rely on textual information or external knowledge.
- **RECK(KG only):** This model focuses solely on using external knowledge from both image and text to perform relation classification. It is equivalent to solely utilizing the "Cross-modal Knowledge Semantic Representation" module in Figure 4.2, while ignoring the "Multimodal Semantic Representation" module. Additionally, the GAT module is not used in this model.
- **RECK(text+image):** To enhance the RECK model that solely relies on textual input, this model incorporates visual features. Specifically, it only utilizes the "Multimodal Semantic Representation" module in Figure 4.2 to perform relation classification.
- **RECK(text+KG):** This model extends RECK(text only) by incorporating knowledge features. It is identical to the RECK model with a single modality, as described in section IV-C.

- **RECK(text+image+KG):** This model adds external knowledge semantic features to RECK(text+image). Compared to RECK, it does not use GAT for further deep feature extraction. Instead, it just feeds them into full connection neural network layers to get advanced features.
- **RECK(text+image+$KG_{v,t}$+GAT):** RECK model without incorporating features from the knowledge graph, which is constructed from the knowledge paths connecting textual seed concepts and visual seed concepts.
- **RECK(text+image+$KG_{t,v}$+GAT):** RECK without incorporating features from the knowledge graph constructed based on the paths from visual seed concepts to textual seed concepts.

Besides, in order to verify the influence provided by the cross-modal knowledge more comprehensively, we set up three kinds of ablation experiments from the following perspectives: (i) using different number of visual seed concepts, (ii) using different source of textual seed concepts, and (iii) using different number of extracted paths between one seed-concept pair. Firstly, we successively extract m visual seed concepts for ablation experiments, where $m = 1, 3, 5, 7, 9$. Secondly, different from the proposed model, we take all the words from the text input except for the stop words as the textual seed concepts to compare the performance of the model. Thirdly, we also choose the shortest n paths for ablation experiments, where $n = 1, 2, 3, 4, 5$, successively.

4.5 Results and Discussion

4.5.1 Statistics of Introduced Knowledge

In total, we have incorporated 17561 cross-modal concepts from ConceptNet over all the instances. In order to demonstrate the diversity of concepts in image posts,

Table 4.2: The Overall Performance of Our Models and Other State-of-the-art Methods (Acc.: Accuracy, Prec.: Precision). * Indicates that the Difference Against F1 Value Between Our Model and Previous Baselines is Statistically Significant By One-Tailed Paired t-test with $p < 0.01$.

Modality	Model	Acc.	Prec.	Rec.	F1.
Text Only	GloVe+CNN (Nguyen et al. 2015) [82]	70.32	57.81	46.25	51.39
	PCNN (Zeng et al. 2015) [117]	72.67	62.85	49.69	55.49
	MTB (Soares et al. 2019) [93]	72.73	64.46	57.81	60.96
Text+Image	VisualBERT(Li et al. 2019) [61]	-	56.34	58.28	57.29
	ViLT (Kim et al. 2021) [53]	73.30	62.80	52.50	57.19
	MTB+SG (Zheng et al. 2021) [128]	74.09	62.95	62.65	62.80
	MEGA (Zheng et al. 2021) [128]	76.15	64.51	68.44	66.41
	HVPNeT (Chen et al. 2022) [17]	-	83.64	80.78	81.85
	MKGformer (Chen et al. 2022) [16]	-	82.67	81.25	81.95
	RECK (Our method)	**95.11**	**88.77**	**88.91**	**88.84***

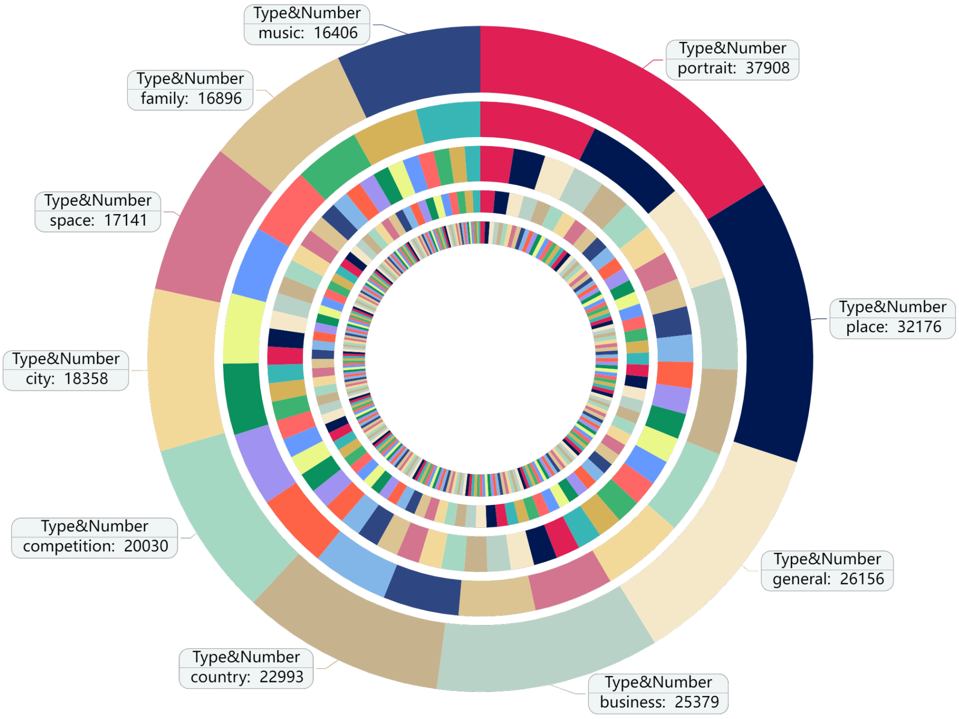


Figure 4.4: Analysis of the introduced concept distribution. We show a few of selected concepts with the top numbers of proportion. Different colors are used to indicate different concepts.

we show the concept distributions in Figure 4.4. As can be seen there, our dataset involves a variety of topics such as portrait, business, competition, and city, etc. We observe that concepts are nearly uniformly distribution which implies the concept balance in our dataset.

Moreover, we visualize the word cloud of top-50 most frequent introduced concepts in Figure 4.5. Among them, both generic (e.g. play) and domain-specific (e.g. business) words are included, and such abundant cross-modal external knowledge surely helps the model to find a path bridging the semantic gap between two modalities.

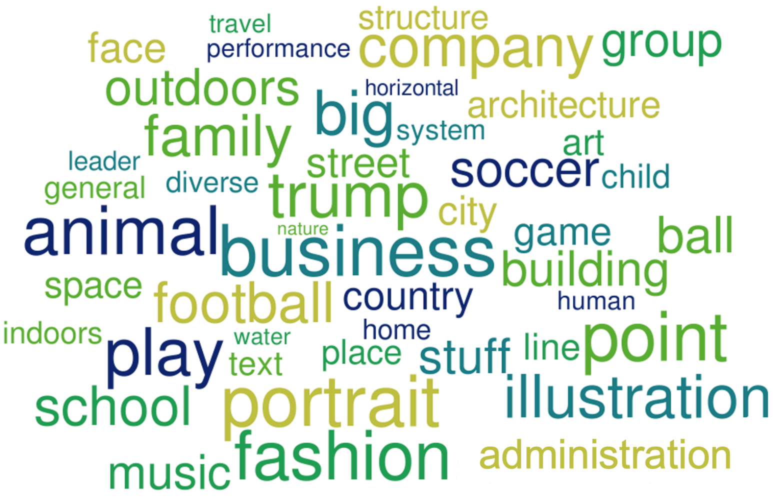


Figure 4.5: Word cloud of concepts extracted from ConceptNet. We show word cloud for top-50 most frequent introduced concepts.

4.5.2 Comparison with Existing Methods

Table 4.2 shows the evaluation results of our RECK model and other existing models on the MNRE dataset from single-modality and multi-modality. There are several findings from the table:

- Firstly, rows 1 to 3 are the results of text based models. Without exploiting the visual information, they suffer from the loss of contexts caused by the extremely short text, leading to unsatisfactory performance.
- Secondly, compared to the previous multimodal models (in rows 4 to 9) which ignore the modality semantic gap and implicitly fuse the information of the two modalities, our method exploits the knowledge paths to bridge the semantic gap and connects the two modalities explicitly via external cross-modal knowledge. The model performance improves because of bridging the modality semantic gap and the more efficient use of multimodal information. As a result, RECK significantly outperforms the state-of-the-art method MKGformer in all metrics, increasing the precision by 6.10%, recall by 7.66% and F1 score by 6.89%.

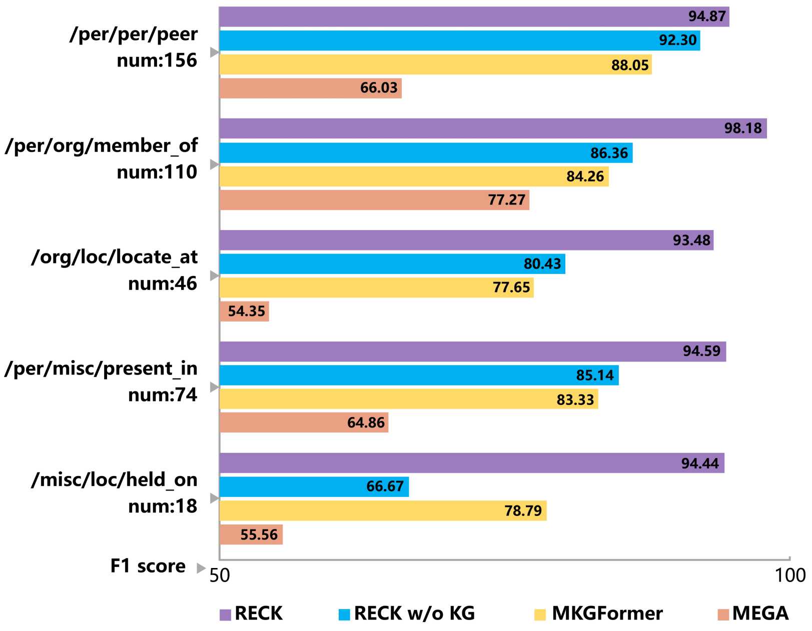


Figure 4.6: The experimental results of five specific categories in the dataset. The comparison models include: RECK, RECK w/o KG, MKGformer, MEGA.

- Thirdly, we find that multimodal pre-trained model like VisualBert and ViLT do not achieve great performance in the MNRE task. Compared to the text based pre-trained model MTB, they even obtain worse scores in most of the metrics. We think the reason might be due to the great semantic gap between textual entities and visual patches, making it hard to capture the modality interaction. Another reason may be that the pre-training datasets they used are quite different from MNRE.
- Fourthly, HVPNet and MKGformer adopt a different method to extract visual features compared to other multimodal models and achieve a huge improvement in performance. However, they just focus on how to better extract visual information, ignoring the semantic differences between images and texts. In contrast, for multimodal tasks, our model can improve the performance not only by mining the information of images and texts, but also by mining the explicit relations or connections between them.

4.5.3 Performance Analysis of Different Categories

In this subsection, we select five categories based on the number distribution from the dataset for specific performance analysis.

They are "/per/per/peer", "/per/org/member_of", "/org/loc/locate_in", " /per /misc /present _in " and "/misc/loc/held_on". Unlike "/per/per/family", "/per /per /couple" and other relations which are easily reflected in visual objects from image or words from text, relations like "/misc/loc/held_on/" require deeper understanding about the text-image pairs. Here, we choose RECK, RECK w/o KG, MKGformer and MEGA for performance comparison.

As shown in Figure 4.6, our model performs the best in all the mentioned categories. Especially in the last four categories, RECK has obvious advantages. We think the incorporated external knowledge plays an important role for the improved performance. For example, to identify the relation "present_in", we need to figure out that the scenery is related to "activity" or "show", other than just rely on the extracted objects "person" or "light". However, the previous methods fail to do well because they ignore the semantic gap between modalities. In comparison, RECK still perform well even on such abstract categories.

4.5.4 Analysis of Computation Cost

As we incorporate external knowledge into our model, efficiency may become a significant concern. Hence, we present a computational cost analysis of our methods (RECK and its variants) compared to multimodal RE baselines within Figure 4.7. We measure computational cost by the time it takes to classify all instances in the test set. For the sake of accuracy, the time mentioned is the average duration needed for conducting test set inference ten times. Firstly, we illustrate the time cost of visual information and external knowledge by removing image representation and

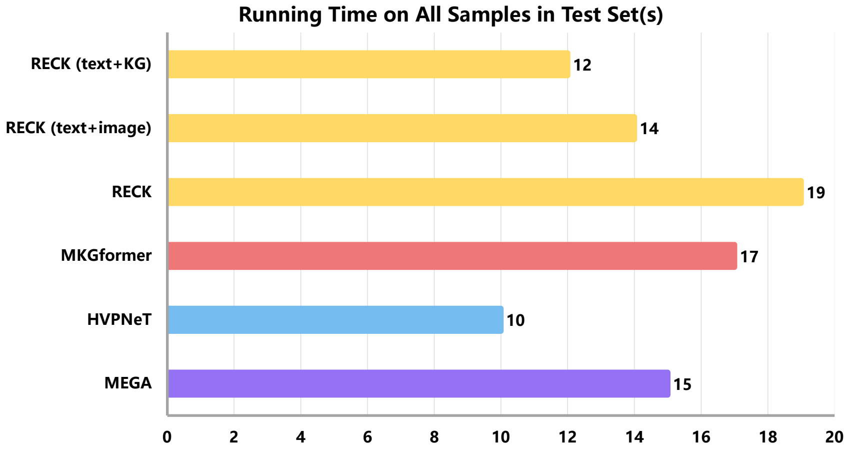


Figure 4.7: Anlysis of computation cost between our proposed model RECK with other SOTA models. Under the same experimental conditions, we use the running time comparison on all the samples from test set to show the computation cost differences between RECK and other models.

knowledge graphs. The outcomes indicate that incorporating visual and knowledge information leads to an increase in computational cost by 1.58 and 1.36 times. It also suggests that visual information imposes a greater computational burden than external knowledge. Secondly, RECK is comparable in computation cost to other SOTA models, which consist of multi-layer transformers or complex graph structural alignment, such as MEGA and MKGformer. And in Table 4.2, RECK outperforms these two models with F1 scores improvement of 22.43% and 6.89%, respectively. Therefore, we think the increase in computational cost is acceptable compared to the improvement in F1 scores.

4.5.5 Ablation Study

We now conduct several ablation studies on the proposed method to illustrate the contribution of each model component.

Table 4.3: The ablation experiment results of RECK variants. BOLD: THE BEST PERFORMANCE IN THE COLUMN. KG means knowledge graph composed of knowledge paths.

Method	Acc.	Prec.	Recall	F1.
RECK(text only)	75.34	64.40	63.59	63.99
RECK(image only)	73.17	33.60	32.34	32.96
RECK(KG only)	62.27	51.25	32.03	39.42
RECK(text+image)	92.01	82.00	81.09	81.54
RECK(text+KG)	78.87	69.07	71.88	70.44
RECK(text+image+KG)	93.99	86.54	86.41	86.47
RECK(text+image+$KG_{v,t}$+GAT)	94.67	88.09	87.81	87.95
RECK(text+image+$KG_{t,v}$+GAT)	94.49	87.85	87.03	87.44
RECK	**95.11**	**88.77**	**88.91**	**88.84**

Effect of Model Components

We first discuss the effect of the model components in RECK. As shown in Table 4.3, all the components in our model play important roles in improving the performance. Based on the results, we have the following findings:

- To begin with, when comparing the first, fourth, and last rows, it is evident that the model's performance improves significantly as visual and external knowledge information is sequentially added, in contrast to the text-only model. As per our initial aim, RECK outperforms RECK(text+image) by a considerable margin. The inclusion of external knowledge enhances RECK's F1 score from 81.54 to 88.84. It is indeed compelling that knowledge paths can effectively bridge the semantic gap between the modalities by integrating them, enabling the model to comprehend multimodal information more effectively.
- Next, we proceeded to segment the model's three primary inputs to execute the

relation classification task individually. The outcomes of the three inputs are presented in the first, second, and third rows, respectively. The results indicate that the use of textual features alone yields the best performance, with an F1 score of 63.99, followed by knowledge features alone and visual features alone. We believe this is reasonable because the relationship between entities is heavily reliant on the text's syntactic and semantic information for reasoning. In comparison to text, visual or knowledge features do not contain robust semantic information. Moreover, since knowledge is more closely related to textual semantics than visual semantics, RECK(KG only) outperforms RECK(image only).

- Thirdly, upon comparing the fourth and fifth rows, we observe that RECK(text+ image) outperforms RECK(text+KG) despite RECK(KG only) surpassing RECK(image only). Our speculation is that this is due to the cross-modal attention mechanism, which facilitates the interaction and alignment of textual and visual features, whereas textual and knowledge features are merely concatenated. The former approach proves to be more effective in integrating cross-modal information than the latter.
- Fourthly, we can draw a conclusion by comparing the fourth, sixth, and last rows that even when GAT is not used, the introduced external knowledge has a beneficial effect on relation classification, with the F1 score increasing from 81.54 to 86.47. By comparing the fifth and last rows, we can observe that the use of GAT further enhances the F1 score from 86.47 to 88.84. This indicates that GAT has a positive impact on capturing underlying semantic dependencies and extracting crucial information from external knowledge by adjusting the concept feature representation.
- Finally, comparing the seventh, eighth, and last rows, we can see that both cross-modality knowledge-aware graphs $KG_{t,v}$ and $KG_{v,t}$ play a crucial role in bridging semantic level discrepancies. While the performance improvement is similar when each graph is used alone, using $KG_{v,t}$ alone has a slightly better

Table 4.4: The results of ablation study based on different number of visual seed concepts. Top-m denotes that there are m seed concepts with highest confidence introduced into the our model for knowledge path retrieval.

Visual Seed Concepts Num	Acc.	Prec.	Recall	F1.
RECK(text+image)	92.01	82.00	81.09	81.54
+ Top-1	94.30	87.78	86.41	87.09
+ Top-3	94.80	89.08	87.97	88.52
+ Top-5	**95.11**	88.77	88.91	**88.84**
+ Top-7	94.80	**89.30**	87.34	88.31
+ Top-10	92.81	83.46	**89.06**	86.17

effect. When both graphs are used together in the model, the performance increases to 88.84, compared to 87.95 and 87.44 when only using one of them. This indicates that constructing knowledge paths from different directions in ConceptNet can help fill the semantic gap from various perspectives. This two-way approach can introduce more concepts at a high semantic level and exploit more comprehensive information to fill the semantic gap.

Analysis of External Knowledge Extraction

After discussing the impact of different modules of RECK, we consider the influence of the method of knowledge path extraction. On the one hand, introducing too much external knowledge will bring noise; on the other hand, presenting only a small part of external knowledge will lead to insufficient use of auxiliary information and fail to bridge the semantic gap. Therefore, to explore the appropriate way to introduce knowledge, we set up three kinds of experiments based on (i) the different number of visual seed concepts, (ii) different sources of textual seed concepts, and (iii) the

Table 4.5: The results of ablation study based on different source of textual seed concepts. RECK *ALL* means all the non stop words from text input are used as textual seed concepts, which is different from RECK itself.

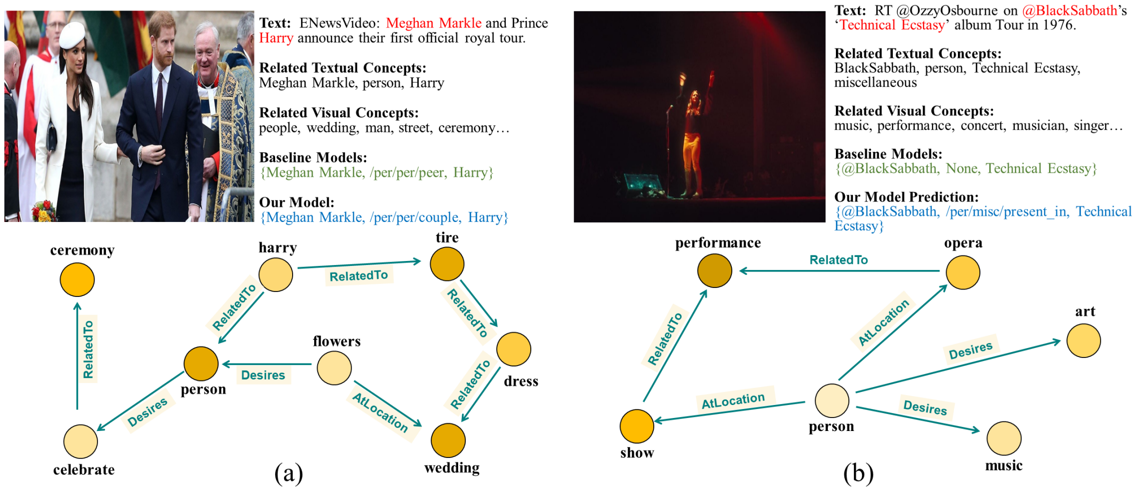
Textual Seed Concept Resource	Acc.	Prec.	Recall	F1.
RECK(text+image)	92.01	82.00	81.09	81.54
RECK *ALL*	93.25	84.05	84.84	84.45
RECK	**95.11**	**88.77**	**88.91**	**88.84**

different number of knowledge paths between each paired text-image seed concepts.

As shown in Table 4.4, the model performance increases gradually when the number of the introduced concepts increases within an appropriate range. We find that the model performs the best when the top-5 visual seed concepts are in use. The visual seed concepts with low confidence may bring some noise knowledge paths irrelevant to the textual information. However, the F1 score drops significantly when we take top-10 seed concepts into account, compared with the one with top-5 seed concepts (from 88.84% to 86.17%).

Besides, when the number of visual seed concepts is 3, 5, or 7, we find that the model's performance is similar. This may be because the extracted visual seed concepts are similar in these cases, resulting in a large number of coincident nodes in each path. Therefore, the constructed knowledge graph becomes similar to each other.

As shown in Table 4.5, when we use all non-stop words in the text as the textual seed concepts, taking F1 as an example, the performance decreases from 88.84 to 84.45. This indicates that although taking more words as seed concepts can introduce more external knowledge, keeping doing so will also lead to more noise, which will degrade the model's performance. Not all the words in a sentence are related to the



Text: ENewsVideo: Meghan Markle and Prince Harry announce their first official royal tour.

Related Textual Concepts:
Meghan Markle, person, Harry

Related Visual Concepts:
people, wedding, man, street, ceremony…

Baseline Models:
{Meghan Markle, /per/per/peer, Harry}

Our Model:
{Meghan Markle, /per/per/couple, Harry}

Text: RT @OzzyOsbourne on @BlackSabbath's 'Technical Ecstasy' album Tour in 1976.

Related Textual Concepts:
BlackSabbath, person, Technical Ecstasy, miscellaneous

Related Visual Concepts:
music, performance, concert, musician, singer…

Baseline Models:
{@BlackSabbath, None, Technical Ecstasy}

Our Model Prediction:
{@BlackSabbath, /per/misc/present_in, Technical Ecstasy}

Figure 4.8: Two examples for illustrating the effectiveness of incorporating cross-modal knowledge representations and the graph attention mechanism. We export subgraphs of external knowledge graphs for showing the related cross-modal concepts, and visualize the attention weights of graph attention network, where the deeper the color, the higher the attention weights.

relationship between entities. This also hints that when using external knowledge to connect multimodal information, the concepts of indexing knowledge need to be designed independently for different multimodal tasks.

As shown in Table 4.6, the overall results achieve the best when the number of knowledge paths is 3. The performance also increases gradually when the number of introduced concepts increases within an appropriate range. However, compared to Table 4.4, the model performance based on the different number of paths is similar. This is because the shortest n paths connecting two concepts in ConceptNet are generally similar, having only a few nodes different between paths, thereby introducing similar knowledge.

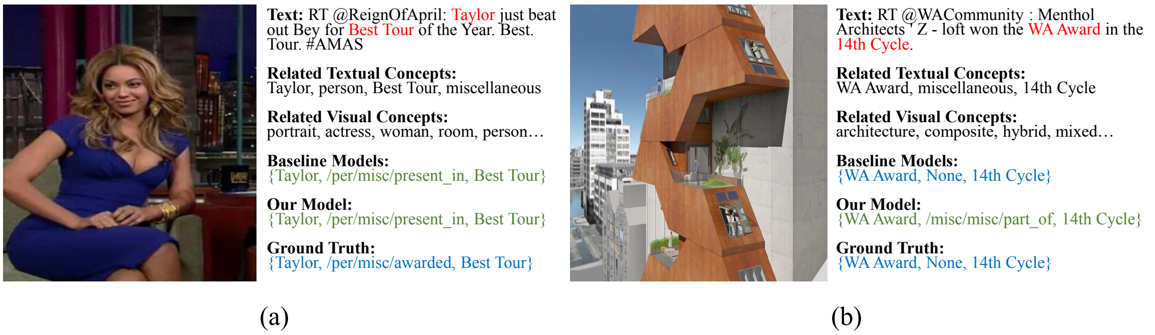
Table 4.6: The results of ablation study based on different number of knowledge paths connected by each textual and visual seed concept pair. Shortest-n denotes that there are n shortest knowledge paths between each seed concept pair introduced into the our model for knowledge path retrieval.

Knowledge Paths Num	Acc.	Prec.	Recall	F1.
RECK(text+image)	92.01	82.00	81.09	81.54
+ Shortest-1	94.73	87.10	88.13	87.92
+ Shortest-2	94.73	88.63	87.66	88.14
+ Shortest-3	**95.11**	88.77	**88.91**	**88.84**
+ Shortest-4	94.80	**88.99**	87.19	88.08
+ Shortest-5	94.67	88.05	87.50	87.78

4.5.6 Case Study

As a case study, we compare our model with baseline methods which leverage the low-level visual object information while ignoring the semantic gap of visual objects and textual relations. In Figure 4.8(a), the example depicts that previous methods incorrectly identify the relation between “Meghan Markle” and “Harry” as “peer”, due to that they ignore the visual information related to concepts “wedding” or “ceremony”. In contrast, our model makes the right prediction with the guidance of these high-level concepts. Moreover, we can see that our graph attention network gives higher weights (i.e., with darker colors) on visual concepts which are more related to the ground-truth relation “Couple”.

The example in Figure 4.8(b) also reveals that previous methods cannot figure out the relation “present in” with only a person extracted as supplementary object information. Nevertheless, we can learn from the knowledge graph that there is a person related to art or music (showing that she may be a singer), appearing on



Text: RT @ReignOfApril: Taylor just beat out Bey for Best Tour of the Year. Best. Tour. #AMAS

Related Textual Concepts:
Taylor, person, Best Tour, miscellaneous

Related Visual Concepts:
portrait, actress, woman, room, person…

Baseline Models:
{Taylor, /per/misc/present_in, Best Tour}

Our Model:
{Taylor, /per/misc/present_in, Best Tour}

Ground Truth:
{Taylor, /per/misc/awarded, Best Tour}

(a)

Text: RT @WACommunity : Menthol Architects ' Z - loft won the WA Award in the 14th Cycle.

Related Textual Concepts:
WA Award, miscellaneous, 14th Cycle

Related Visual Concepts:
architecture, composite, hybrid, mixed…

Baseline Models:
{WA Award, None, 14th Cycle}

Our Model:
{WA Award, /misc/misc/part_of, 14th Cycle}

Ground Truth:
{WA Award, None, 14th Cycle}

(b)

Figure 4.9: An error analysis was conducted to compare our proposed RECK model with the baseline method. Images are displayed on the left side, while the detected textual and visual concepts, along with the predictions, are presented on the right side. Two specific scenarios were examined in these cases: 1) both the baseline model and RECK made incorrect predictions, and 2) the baseline model produced the correct prediction while RECK failed.

a show/opera which is related to performance. Such auxiliary information strong indicates the relation between "BlackSabbath" and "Technical Ecstasy" as "present in", rather than "None" as indicated by other compared baseline methods.

4.5.7 Error Analysis

We also show some examples failed by RECK and previous baseline methods in Figure 4.9. The first case in Figure 4.9(a) reveals that when both the text and image can not provide enough information to identify the relation, our model may output an incorrect result based on limited contents. In this example, our model extracts the relation "present in" between "Taylor" and "Best Tour" with the auxiliary visual information "a person in the room". However, it cannot find any clues to indicate the ground-truth label "awarded".

The second example in Figure 4.9(b) presents another case that external knowledge has the opposite effect on relation extraction. Our model incorrectly predicts the relation "part_of" since some related concepts (e.g., composite, hybrid and mixed) are extracted, although the two entities show no relation under the predefined categories. On the contrary, baseline methods without incorporated knowledge can output the ground truth label.

4.6 Conclusion

In this work, we propose a novel model called RECK for multimodal knowledge extraction. Our model leverages cross-modal knowledge representation to bridge the semantic gaps between visual contents and textual relations of entity pairs. We first extract seed concepts from text and image input and then retrieve knowledge paths from external knowledge base. In addition, we adopt GAT module to filter the knowledge and generate the cross-modal knowledge representation.

The experimental results on the MNRE dataset demonstrate that our model outperforms the state-of-the-art methods by successfully bridging the gap. For future work, we plan to develop and devise more efficient and explainable filtering methods and path selection strategies.

Chapter 5

A Robust Data Augmentation and Estimation System for Knowledge Extraction

5.1 Introduction

Multimodal language understanding has received intensive attention recently for its advantage of mining semantics by collaborating the cross-modal inference [113]. Examples include methods for multimodal name entity recognition (MNER) [123] and multimodal relation extraction (MRE) [128]. Both benefit from the collaborative reasoning based on the alignment of textual and visual content. However, statistics on commonly adopted text-image relation benchmarks (e.g., TRC [101] and Twitter100k [45]) shows that the misalignment rate between images and texts is as high as 60%. Noise introduced by the misalignment can mislead the learning and degrade the performance of resulting models.

As shown in Fig. 5.1, the misalignment can be categorized into *partial* and *irrele-*

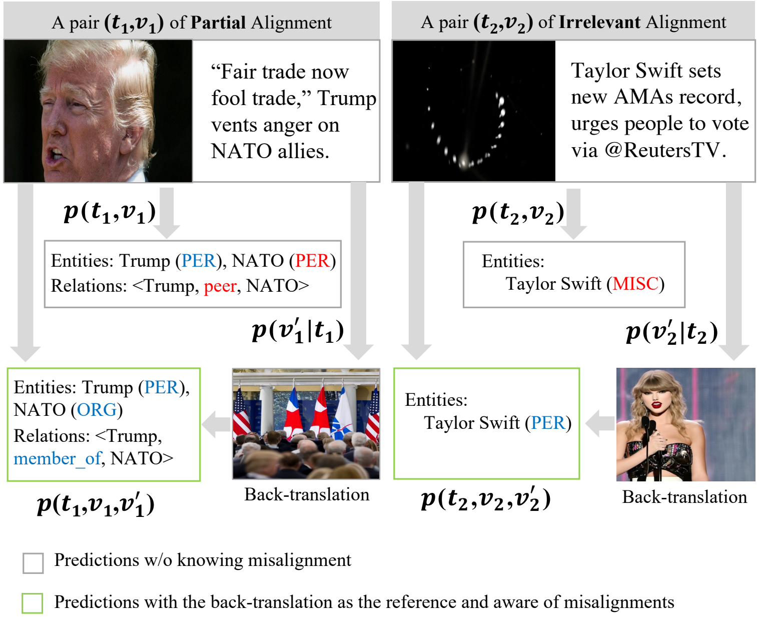

A pair (t_1,v_1) of **Partial** Alignment
"Fair trade now fool trade," Trump vents anger on NATO allies.
$p(t_1,v_1)$
Entities: Trump (PER), NATO (PER)
Relations: <Trump, peer, NATO>
$p(v_1'|t_1)$
Entities: Trump (PER), NATO (ORG)
Relations: <Trump, member_of, NATO>
$p(t_1,v_1,v_1')$
Back-translation
A pair (t_2,v_2) of **Irrelevant** Alignment
Taylor Swift sets new AMAs record, urges people to vote via @ReutersTV.
$p(t_2,v_2)$
Entities: Taylor Swift (MISC)
$p(v_2'|t_2)$
Entities: Taylor Swift (PER)
$p(t_2,v_2,v_2')$
Back-translation
□ Predictions w/o knowing misalignment
□ Predictions with the back-translation as the reference and aware of misalignments

Figure 5.1: Partial (left) and irrelevant (right) alignments in text-image pairs and the results of using generative back-translation to help the inference in multimodal entity and relation extraction tasks.

vant alignment. In case of incomplete alignment, textual entities (e.g., NATO) might be mismatched to the visual evidence (e.g., person) which results in incorrect labels (e.g., PER). This further leads to underline relations between entities (e.g., <Trump, president of, USA), <USA, member of, NATO>) missing from the extractions. In case of irrelevant alignment, the textual entities might be randomly matched to visual evidence (e.g., MISC) resulting in dirty data for inference. While the misalignment with the ambiguity/distraction it brings to the learning has long been noticed, it has been rarely studied and addressed [96]. The challenge is that it is nearly impossible to know the degree of misalignment prior to the inference. Otherwise, the inference may has already been done.

In this work, we conduct a pilot study to address this problem. The motivation is that the misalignment of cross-modal pairs is a similar problem to the divergence of cross-lingual machine translations [12]. The problem can thus be transformed by treating the text-image pairs in MNER/MRE as translations to each other. The divergence problem is more widely studied and existing solutions such as back-translation [25] can be borrowed.

While this sounds appealing, it introduces new challenges as follows.

Modality Gap: The cross-lingual divergence is defined in a monomodal setting. The divergence can be measured explicitly by using features such as difference of sentence lengths, ratio of aligned words, and number of unaligned contiguous sequences [12]. However, those features are not available in a cross-modal setting. We address it in an implicitly way in which disalignment of cross-lingual words (e.g., textual words and visual patches) is indicated by the divergence of their representations in the embedded space.

Parallelism: The detection/assessment of cross-lingual divergence relies on large-scale parallel corpora, in which the sentences are aligned into word-level. The alignment is symmetric which makes high quality back-translation possible. However, in the cross-modal setting, MNER/MRE benchmark datasets are with a small scale due to the high cost of name entities labeling. The datasets are not well paralleled and there is no word-level alignment. We address those problems by taking advantage of the latest development of diffusion-based generative models [89]. Those models are trained on large-scale and better paralleled datasets, with which the back-translation can be conducted in a generate-to-translate way, in a sense that, for each text sentence, we can generate an image as its visual language "translation". Visual grounding [114] can then be employed to make the alignment into word-level. More details will be given in Section 5.3.2.

Low-Resource Benchmarks: The assessment of the divergence needs datasets on large-scale. This is not the case in MNER/MRE scenario. We borrow the idea of using high-resource corpora as a bridge to address the low-resource learning issue [32, 28]. In this papaer, a new multimodal dataset is constructed for multimodal diver-

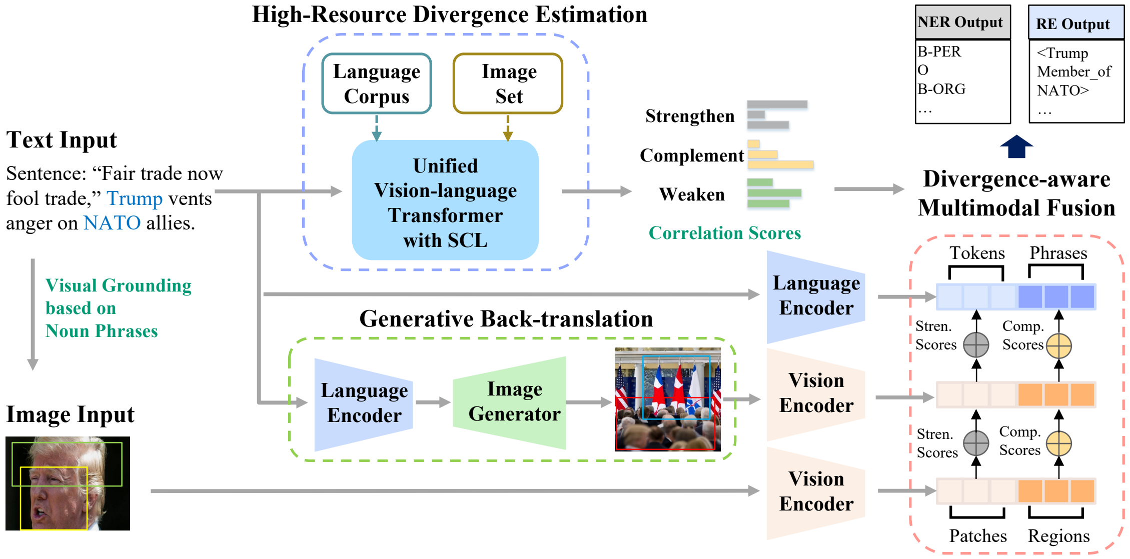


Figure 5.2: The framework of the proposed **T**ranslation motivated **M**ultimodal **R**epresentation learning (TMR), which generates divergence-aware cross-modal representations by introducing two additional streams of Generative Back-translation and High-Resource Divergence Estimation.

gence estimation. An estimator is built which generates fine-grained confidence scores over 3 alignment categories of *strengthen, weaken, and complement*. It enables better argumentation for MNER/MRE than the simple similarity-based filtering schemes adopted previously. It also preserves the text-image pairs that are not well-aligned but with complementary evidence. More details will be given in Section 5.3.3.

5.2 Problem Statement

Definition 1. ***(Multimodal Translation)*** *Give a pair of a sentence t and an image v, our interest is the joint probability $p(t, v)$, on the basis which the "translation" using either modality as the source "language" can be obtained/evaluated (e.g., using $p(t \mid v)$ or $p(v \mid t)$) [12].*

However, in the multimodal information extraction scenario, the translation is not a goal. We use it as a conceptual solution-seeking mindset. Specifically, our target is to build a function $g(t, v)$ which learns the representations of $p(t, v)$. We propose to make the learner aware of the modality misalignment (divergence) using

- Back-Translation: a generative diffusion model is employed as a predictor for $p(v' \mid t)$ which generates the back-translation of v. The divergence can be embedded by integrating the representations of v and v';
- High-Resource Divergence Estimation: we learn a function $d(t, v)$ to estimate the cross-modal divergence. The function is learned on a high-resource corpora independently and can be used to adjust $p(t, v)$.

5.3 The Proposed System: TMR

In this section, we introduce a general process for learning the representation first (i.e., $g(t, v)$), and then $p(v' \mid t)$ and $d(t, v)$ can be implemented. Once the representation is obtained, multimodal information extraction tasks such as NER and MNRE can be conducted by learning the probability of $p(l \mid g(t, v))$ where l represents the label of name entities or relations depending on the task. The framework is shown in Fig. 5.2.

5.3.1 Multi-Grained Representation Learning

To ease the description, let us denote the resulting representation of a text-image pair as $\boldsymbol{\mathcal{G}} = g(t, v)$. It can be implemented using a Transformer model [51] as long as t and v can be tokenized (e.g., into words or patches) and embedded, so that the joint representation is learned regarding the cross-model correlation (ensured by the multi-head attention). Denote $\boldsymbol{T}$ and $\boldsymbol{V}$ as the tokenized embedding of t and v, respectively, the representation can be learned as

$$\boldsymbol{\mathcal{G}} = \sum softmax\left(\frac{\boldsymbol{W}_d \boldsymbol{V} \boldsymbol{T}^\top}{\sqrt{d}}\right)\boldsymbol{T}, \tag{5.1}$$

where d is the dimension of textual embedding $\boldsymbol{T}$ and $\boldsymbol{W}_d$ is a cross-model attention matrix which is learned during the training.

However, granularity is a concern when the representation is cross-modal, because of the aforementioned Modality Gap and Parallelism challenges. We propose to build a multi-grained representation learning scheme, in which a 2-level of granularity is adopted so that a text is tokenized into words and phrases and an image is tokenized into patches and regions. We assume that the cross-modal representation can be generated on a fine scale based on word-patch correlations and the representation is coarse-grained when built on phrase-region correlations [63].

Let us denote $\boldsymbol{T}^w$ and $\boldsymbol{T}^p$ as the tokenized embedding of the text t at word and phrase level, respectively, in which the phrases is obtained using Stanford Parser following the method in [119]. The embedding are encoded using BERT [51]. Similarly, we denote $\boldsymbol{V}^s$ and $\boldsymbol{V}^r$ as the tokenized embedding of the image v at patch and region level, respectively, in which patches are obtained using fixed grid and regions are obtained using the visual grounding method toolkit [114]. We set the numbers of patches and regions as 49 and 3, respectively, by following the previous studies [17, 16]. ResNet50 [38] is then employed to generate the visual embedding. The 2 levels of pairs $(\boldsymbol{T}^w, \boldsymbol{V}^s)$ and $(\boldsymbol{T}^p, \boldsymbol{V}^r)$ are then be substituted into Eq. (5.1), resulting in the cross-modal representations $\boldsymbol{\mathcal{G}}^f$ and $\boldsymbol{\mathcal{G}}^c$ at fine and coarse level, respectively. A multi-grained representation $\boldsymbol{\mathcal{G}}$ can then be generated as

$$\boldsymbol{\mathcal{G}} = \boldsymbol{\mathcal{G}}^f + \boldsymbol{\mathcal{G}}^c. \tag{5.2}$$

5.3.2 Cross-Modal Back-Translation

We borrow the idea of back-translation from traditional machine translation methods [25], in which the result in the target language is translated back to the source language to verify the quality or divergence. In our case, we treat the text t as a translation from an image v. A back-translation v' can then obtained by using

$$v' = \arg\max\ p(\hat{v} \mid t), \tag{5.3}$$

where $\hat{v}$ is an image hypothesis. However, back-translation usually requires parallel corpora to learn the probability of $p(\hat{v} \mid t)$, which is not available in any NER/MNRE settings. We address this problem by taking advantage of recent advance in diffusion-based generative models [89]. Those models are trained using large-scale paralleled text-image pairs to learn the ability to generate an image contained on a give text prompt. The objective of those models is thus conceptually similar to Eq. (5.3). In our case, we use stable diffusion [88], which is trained on a subset of LAION-5B [90] dataset. Upon back-translation, we feed the text t as a prompt to stable diffusion. The modal generates a v' which can be used as an approximation of the back-translation from t.

To assess the divergence of translation, we cannot compare v' to v like in text translation, because the cross-modal misalignment is at the semantic level and indicated by the correlation rather than the content. We thus compose a new pair (t, v') and use the process introduced in Section5.3.1 to generate a back-translated cross-modal representation $\boldsymbol{\mathcal{G}}'$. Since v' is generated directly from t, the alignment between them is better guaranteed than those sampled from user generated content on web or social media. It can be used a pseudo-paralleled pair. Therefore, the original pair (t, v) is better aligned if $\boldsymbol{\mathcal{G}}$ is similar to $\boldsymbol{\mathcal{G}}'$ or otherwise less aligned. There are different ways to use these two representations complementarily. Examples will be given in Section 5.3.4 under MNER/MRE scenario.

5.3.3 High-Resource Divergence Estimation

In this subsection, we implement an independent divergence estimator $d(t, v)$. Existing methods address the issue by setting an attention mask on the reasoner trained on low-resource NER/MNRE benchmarks which simply filters out the less attended pairs [123, 110]. We argue that the training is easy to be biased by replying low-resource benchmarks which are neither sufficient on scale nor designed for divergence assessment purpose. More importantly, the filtering scheme also ignores pairs that are less aligned but with complementary evidence (e.g., Fig. 5.1). We construct a high-resource corpora which serves as a bridge to train the estimator independently. Furthermore, the estimator generates for each pair 3 confidence scores $(\alpha_s, \alpha_c, \alpha_w)$ over the category set {*strengthen, complement, weaken*} for a more detailed divergence estimation. It can then be utilized as an augmenter (instead of a filter) for better representations of $\boldsymbol{\mathcal{G}}$ and $\boldsymbol{\mathcal{G}}'$ as

$$\begin{bmatrix} \boldsymbol{\mathcal{G}}^* \\ \boldsymbol{\mathcal{G}}'^* \end{bmatrix}^\top = \begin{bmatrix} \alpha_s \\ \alpha_c \\ \alpha_w \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\mathcal{G}}^f & \boldsymbol{\mathcal{G}}'^f \\ \boldsymbol{\mathcal{G}}^c & \boldsymbol{\mathcal{G}}'^c \\ 0 & 0 \end{bmatrix}, \qquad (5.4)$$

$$w.r.t. \;\; \alpha_s + \alpha_c + \alpha_w = 1.$$

High-Resource Corpora Construction Different from [96] using limited data crawled from social media (e.g., Twitter), we collect data from large-scale public image-text datasets to enhance the generalization of our estimator. We randomly select 100k data from MSCOCO [66] as the “Strengthen” samples, since the dataset contains fine-grained aligned image-text pairs designed for tasks like Visual grounding and Scene graph generation. LAION-400M [91] is chosen as the “Complement” dataset since it is built on web paired data and no strict rules are applied for the alignment between image contents and text tokens. Similar to MSCOCO, we select 100k image-text pairs from LAION-400M as training samples. We generate negative samples as the “Weaken” (unaligned) data by substituting the images in the

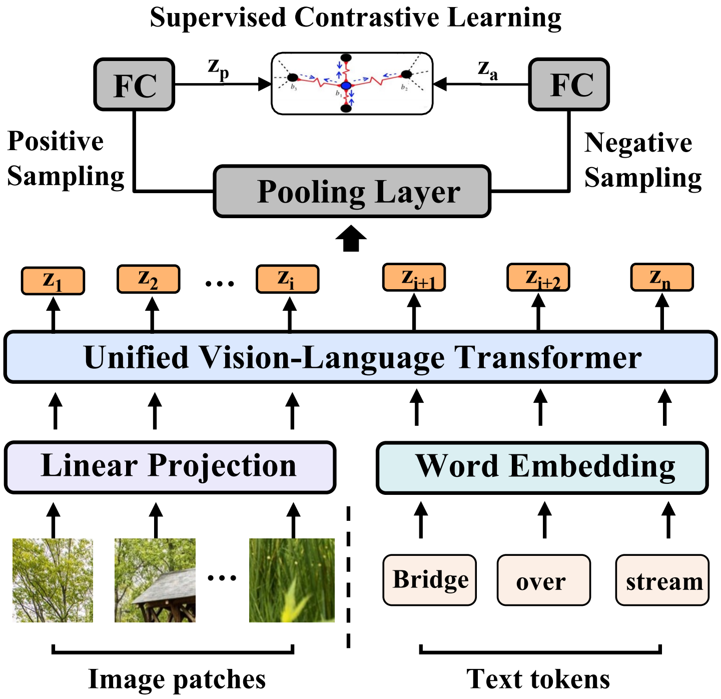


Figure 5.3: Architecture of our Multimodal Divergence Estimator (MDE), which is trained on high-resource vision-language datasets, and Supervised Contrastive Learning (SCL) is applied to enhance the generalization.

“Strengthen” and “Complement” data with a different image randomly sampled from the two datasets. Finally, we accumulate 400k training samples, with 100k, 100k, 200k for “Strengthen”, “Complement” and “Weaken”, respectively. To verify the effectiveness and generalization, we further construct a in-domain test set of 10k data sampled from the two datasets and a out-of-domain test set of 1k data from the SBU dataset which contains both fine-grained and coarse-grained aligned text-image pairs.

Model Design We adopt the same structure as ViLT [53] that leverages a unified transformer to encode visual and textual contents. To be more specific, the input image v (or its back-translation v') is sliced into patches and flattened. Then a linear projection is applied to transfer the visual features to the same dimensions of token embeddings. The text and image embeddings are concatenated into a sequence Z

and iteratively updated through D-dimensional Transformers. We get the pooled representations of the multimodal input sequence M as final output z. Details can be found in Figure 5.3 and Section 5.5.3.

Supervised Contrastive Learning Conventional supervised methods use Cross-entropy Loss to distinguish samples with different classes. However, since our pre-training data are constructed on different datasets, simply applying cross-entropy loss will lead the model to learn a short-cut by utilizing the domain difference other than the semantic alignment. This results in poor generalization performance. To tackle this problem, we propose to use the supervised contrastive learning [52] instead to push away the distance between anchors and negative samples generated from the positive classes "Strengthen" and "Complement".

A self-supervised learning loss can be written

$$L_{self} = -\sum_{i \in I} log \frac{exp(z_i \cdot z_{j(i)}/\tau)}{\sum_{a \in A(i)} exp(z_i \cdot z_a/\tau)} \tag{5.5}$$

where z is the output of our estimator model, τ is a scalar temperature parameter. i, j, a denote the anchor point, positive and negative samples, respectively. We can simply generalize the Eq. (5.5)to incorporate supervision as:

$$L_{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} log \frac{exp(z_i \cdot z_p/\tau)}{\sum_{a \in A(i)} exp(z_i \cdot z_a/\tau)} \tag{5.6}$$

where $P(i)$ is the set of indices of positives and $|P(i)|$ denotes its cardinality.

5.3.4 Multimodal Knowledge Extraction

We use the augmented representations $\boldsymbol{\mathcal{G}}^*$ and $\boldsymbol{\mathcal{G}}'^*$ for two tasks of NER and MNRE.

Named Entity Recognition Following [17, 116], we adopt the CRF decoder to perform the NER task. We fuse the $\boldsymbol{\mathcal{G}}^*$ with its back-translation $\boldsymbol{\mathcal{G}}'^*$ using using multi-head extension [51] and denoted the final representation for a pair (t, v) as

$$\bar{\boldsymbol{\mathcal{G}}} = Multihead(\boldsymbol{\mathcal{G}}^{*}, \boldsymbol{\mathcal{G}}'^{*}) \in \mathbb{R}^{n \times d} \tag{5.7}$$

which consists of the representation of n words from the text t. NER is then a task to predict probabilities of those words over a set of predefined entity labels (e.g., PER, ORG). Let us denote this label set as $\mathcal{L} = \{l\}$. The probabilities are then denoted as $Y = [y] \in \mathbb{R}^{n \times |\mathcal{L}|}$ and calculated as

$$p\big(y \mid \bar{\boldsymbol{\mathcal{G}}}\big) = \frac{\prod_{i=1}^{n} F_i(y_{i-1}, y_i, \bar{\boldsymbol{\mathcal{G}}})}{\sum_{l_j \in \mathcal{L}} \prod_{i=1}^{n} F_i(y_{i-1,j}, y_{i,j}, \bar{\boldsymbol{\mathcal{G}}})}, \tag{5.8}$$

where $y_{i,j}$ denotes the probability of the i^{th} word over the j^{th} label, and F represents potential functions in CRF. We use the maximum conditional likelihood estimation as the loss function

$$L_{ner} = -\sum_{i=1}^{n} log\Big(p(y|\bar{\boldsymbol{\mathcal{G}}})\Big). \tag{5.9}$$

Relation Extraction We merge the representations of textual entities, fine-grained and coarse-grained image-text pairs, as well as noun phrases to predict final relations. For a given pair of entities (e_i, e_j) corresponding to the i^{th} and j^{th} words from t, we generate its representation as

$$\ddot{\boldsymbol{\mathcal{G}}}_{i,j} = \boldsymbol{T}_i \oplus \boldsymbol{T}_j \oplus \mathbf{p} \oplus \mathbf{h} \tag{5.10}$$

where $\boldsymbol{T}_i$ and $\boldsymbol{T}_j$ denote the embeddings of the two entities, respectively, $\oplus$ indicates the concatenation operation, $\mathbf{p}$ denote the summed features of noun phrases in the text t, and $\mathbf{h}$ denotes the summed representation of the text-image pair and its back-translation (i.e., $\mathbf{h} = \boldsymbol{\mathcal{G}}^{*} + \boldsymbol{\mathcal{G}}'^{*}$). We can then aggregate the likelihoods of this representation over a set of relation labels $\mathcal{R} = \{r\}$ as $p(r \mid \ddot{\boldsymbol{\mathcal{G}}}_{i,j}) = softmax(\ddot{\boldsymbol{\mathcal{G}}}_{i,j})$. Finally, we can calculate the RE loss with cross-entropy loss function

$$L_{re} = -\sum_{i=1}^{n} log\Big(p(r \mid \ddot{\boldsymbol{\mathcal{G}}}_{i,j})\Big). \tag{5.11}$$

Modality	**Methods**	*Twitter-2015*			*Twitter-2017*		
		Prec.	Rec.	F1	Prec.	Rec.	F1
Text	CNN-BLSTM-CRF	66.24	68.09	67.15	80.00	78.76	79.37
	HBiLSTM-CRF	70.32	68.05	69.17	82.69	78.16	80.37
	BERT-CRF	69.22	74.59	71.81	83.32	83.57	83.44
	PCNN	-	-	-	-	-	-
	MTB	-	-	-	-	-	-
Text+Image	AdapCoAtt	69.87	74.59	72.15	85.13	83.20	84.10
	OCSGA	74.71	71.21	72.92	-	-	-
	RpBERT	71.15	74.30	72.69	-	-	-
	UMT	71.67	75.23	73.41	85.28	85.34	85.31
	UMGF	74.49	75.21	74.85	86.54	84.50	85.51
	VisualBERT	68.84	71.39	70.09	84.06	85.39	84.72
	MEGA	70.35	74.58	72.35	84.03	84.75	84.39
	HVPNeT	73.87	**76.82**	75.32	85.84	87.93	86.87
	MKGFormer	-	-	-	86.98	88.01	87.49
	TMR w/o BT.	74.99	75.18	75.08	84.89	88.16	86.49
	TMR w/o MDE.	74.70	76.05	75.37	85.53	87.93	86.72
	TMR (our method)	**75.26**	76.49	**75.87**	**88.12**	**88.38**	**88.25***

Table 5.1: The Overall Performance of TMR compared to several baselines on three benchmark datasets for MNER. We show the prediction results of TMR variants (without Back Translation (BP) or Multimodal Divergence Estimation (MDE)) in the bottom rows. * Indicates that the Difference Against F1 Value Between Our Model and Previous Baselines is Statistically Significant By One-Tailed Paired *t*-test with $p < 0.01$.

5.4 Experimental Settings

5.4.1 Datasets and Metrics

We adopt three publicly available datasets for evaluating our proposed method on MNER and MRE, including: 1) **Twitter15** [72] and **Twitter17** [123] are two datasets for MNER, which include user posts on Twitter during 2014-2015 and 2016-2017, respectively. 2) **MNRE** [128] is a manually-annotated dataset for MRE task, where the texts and images are crawled from Twitter and a subset of Twitter15 and Twitter17. We use precision, recall and F1 value as the default evaluation metric and compare such results in the following sections.

5.4.2 Baselines

We compare our method with two groups of state-of-the-art (SOTA) methods as follows.

Text-based Methods: *CNN-BLSTM-CRF* [76], *HBiLSTM-CRF* [55], and *BERT-CRF* [51] are classical sequence-labeling methods which show excellent prediction results on NER in newswire domain. *PCNN* [117] is a distantly-supervised method for relation extraction, leveraging the knowledge from external knowledge base. *MTB* [93] is a SOTA method for many text-based RE tasks.

Previous SOTA Multimodal Approaches: *AdapCoAtt* [123] is the pioneer work that extracts named entities with co-attention mechanism. *RpBERT* [96] explicitly calculates image-text similarities by learning a classifier on Twitter data. *OCSGA* [110], *UMT* [116], *UMGF* [119], and *MEGA* [128] are the NER/RE methods that align fine-grained object features with textual representations with Transformers or Graph Neural Networks. *VisualBERT* [61] is a vision-language pretraining model that can be applied for MNER and MRE tasks. *HVPNet* [17] and *MKGFormer* [16], the latest SOTA for MNER and MRE, which develops a hierarchical structure to learn visual prefix from multiple views.

Modality	Methods	*MNRE*		
		Prec.	Rec.	F1
Text	CNN-BLSTM-CRF	-	-	-
	HBiLSTM-CRF	-	-	-
	BERT-CRF	-	-	-
	PCNN	62.85	49.69	55.49
	MTB	64.46	57.81	60.86
Text+Image	AdapCoAtt	-	-	-
	OCSGA	-	-	-
	RpBERT	-	-	-
	UMT	62.93	63.88	63.46
	UMGF	64.38	66.23	65.29
	VisualBERT	57.15	59.48	58.30
	MEGA	64.51	68.44	66.41
	HVPNeT	83.64	80.78	81.85
	MKGFormer	82.67	81.25	81.95
	TMR w/o BT.	88.13	84.69	86.37
	TMR w/o MDE.	89.45	86.09	87.73
	TMR (our method)	**90.48**	**87.66**	**89.05**

Table 5.2: The Overall Performance of TMR compared to several baselines on three benchmark datasets for MRE. We show the prediction results of TMR variants (without Back Translation (BP) or Multimodal Divergence Estimation (MDE)) in the bottom rows.

5.5 Results and Discussion

5.5.1 Comparison to SOTA

The results are shown in Table 1. It is easy to see our method outperforms other SOTA methods on on all datasets.

When compared to models relying on pure textual information, visual features contribute to the performance gain by 5% on MNER and 20% on MRE. Due to the short and ambiguous characteristics of texts in social media, it is difficult to identify entities and their relations in limited context.

Incorporating multi-grained visual and textual information performs better than relying on object or image level information solely. The SOTA method HVPNeT and our MTR gain better results (88.35% and 86.87% in Twitter-2017 dataset) than UMGF (85.51%) and UMT (85.31%) which align image and text in fine-grained object-level.

Our model outperforms HVPNet and MKGFormer which leverage hierarchical visual representations or powerful vision-language pretraining embeddings, in a relatively large margin (from 82% to 89%) on the MRE task. We observe a more obvious performance improvement on MRE datasets compared to that on MNRE. The difference comes from the different distributions of MRE and MNRE datasets. Our statistics show that the proportion of complementary cases is significantly higher in MRE (51.5%) than in MNRE (15.7%). As mentioned in the work, the proposed back-translation helps the two tasks by providing additional contextual information for inference. This benefits the complementary cases the most because it makes the identification of indirect relationships possible (otherwise, those cases will be considered as misalignments or used incorrectly like in the similarity-based methods).

5.5.2 Ablation Study

In this section, we conduct extensive experiments with the variants of our model to analyze the effectiveness of each component.

Back-translation: We ablate the procedure of generating back-translation images and the results in Table 1 show the component can boost model performance by 1-3% in MNER and MRE. Still, our ablated model gains comparable or superior performance against baselines which demonstrates the effectiveness of back-translation.

Multimodal Divergence Estimation: Compared with similarity-score based method RpBERT, our model shows stronger extraction and generalization performance with 3.18% improvement on Twitter-2015 dataset. Also, our model achieves significant improvements (3% to 7%) over attention-based methods, revealing that TMR can improve conventional NER/RE methods by decomposing the divergence into fine-grained level.

5.5.3 Other Essentials of the Model

Low-resource Performance

We conduct experiments in low-resource scenarios following the setting of [17], by randomly sampling 5% to 50% from original training set. From the results in Figure 4, we can observe: 1) The methods utilizing multi-grained features (HVPNet and TMR) consistently outperform object-level models in MNER (UMGF) and MRE (MEGA). Multi-grained features can provide global and local views and help models infer entities and relations efficiently. 2) Moreover, our proposed TMR model performs better than HVPNet with external knowledge from generative diffusion models, which addresses the information lack problem in low-resource scenarios.

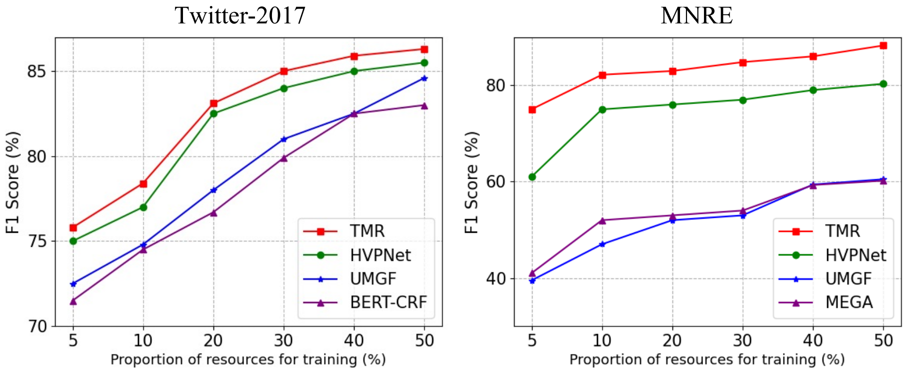


Figure 5.4: Performances in low-resource setting on MNER and MRE tasks.

Improvements on Complementary Cases

To demonstrate the effectiveness of correlation decomposition, we further compare our method with SOTA method HVPNeT on complementary cases of MNRE test set. We argue that previous similarity-based methods ignore the cross-modal divergence, especially when texts and images are complementary. We export 832 cases with "complement score" higher than 0.5 from 1614 test samples. Our model achieves significant improvements against HVPNeT, especially on some categories (e.g., Present in, Locate at and Residence) that rely on deeper understanding of visual scenarios.

Table 5.3: Our results on complementary cases compared to HVPNeT [17] on the MNRE test set. Six main categories are selected for comparison.

Category	Count	TMR	HVPNeT
Peer	98	**91.00**	89.30
Member_of	46	**97.87**	82.11
Contain	33	**98.46**	95.65
Present_in	44	**91.95**	79.01
Locate_at	18	**97.14**	75.68
Residence	13	**83.87**	66.67
Overall	832	**87.37**	77.93

Generalization Performance of Multimodal Divergence Estimator

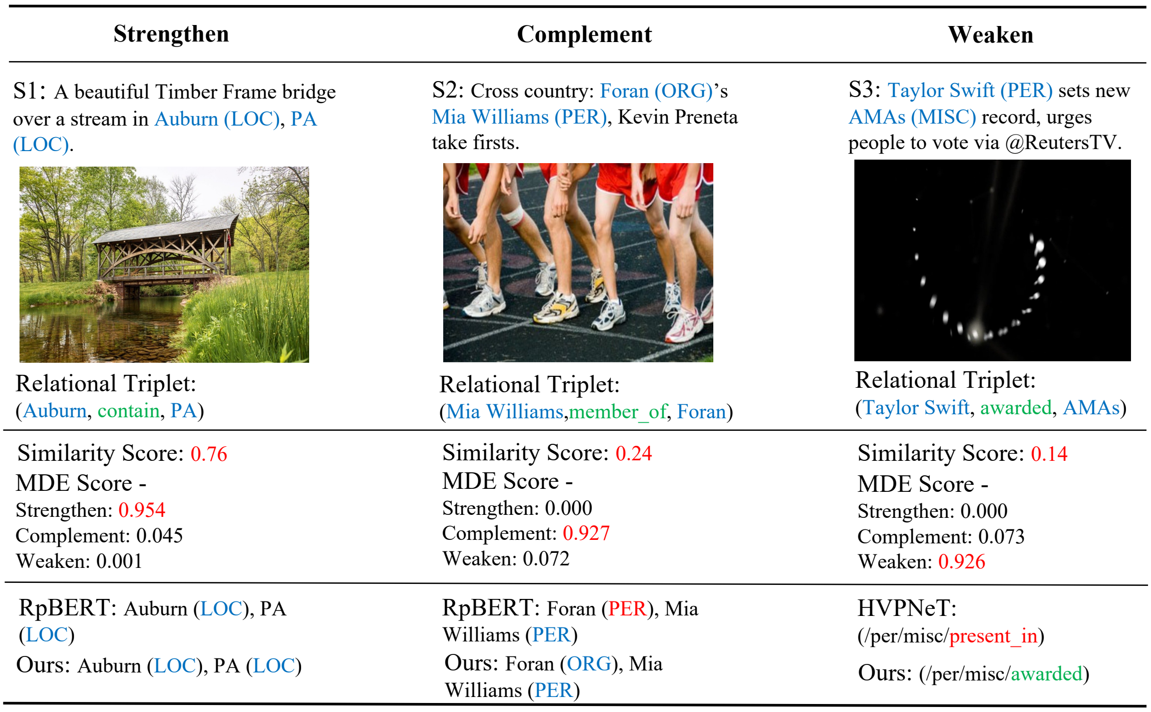
We extend conventional similarity score into fine-grained level and weight the importance of incorporated visual information based on the pretrained divergence estimator. To verify the generalizations to data in other domain, we first construct test set collected with in-domain data (i.e., by sampling on MSCOCO and LAION400M). Then, We first request 2 annotators to label 1k test samples on out-of-domain data and then ask other 2 to review and rectify the test set. As shown in Table 5.4, we compare the estimator trained with different loss function. The results indicate that the model with cross-entropy loss suffers the generalization problem when transferred into out of domain data. The possible reason is that the model may learn a shortcut from the difference of image/text style on the data from the two datasets, other than taking the image-text correlation into consideration. We improve it by introducing negative sampling on in-domain data to reduce the style bias and the F1 value on out-of-domain data increases from 61.8 to 80.01. We further apply the supervised contrastive learning to pull together the positive samples and push apart negative ones, resulting in better generalization performance. The lower in-domain performance but better generalization of supervised contrastive learning can be attributed to: (1) Contrastive learning focuses on learning robust features rather than fitting training distribution, (2) Cross-entropy tends to overfit to training domain-specific features, (3) The learned feature space from contrastive learning better captures semantic similarities.

Model Setting	In Domain	Out of Domain
Cross-entropy	98.56	61.80
Negative Sampling	92.57	80.01
Supervised Contrastive	93.26	**86.21**

Table 5.4: The generalization experiment of the Multimodal Divergence Estimator (MDE). Origin. is the dataset with 10k data sampling from pretraining data, while SBU is the 1k dataset for human evaluation. F1 value is used for evaluation metric.

Case Study

To validate the effectiveness and robustness of our method, we conduct case analysis for multimodal divergence estimation. Previous works simply calculate the image-text similarity with attention mechanism (HVPNeT) or pretrained classifier (RpBERT). As a result, visual information with low similarity score will be filtered out. We notice that our model and RpBERT can identify entities correctly when images are well-aligned with sentence in S1. However, RpBERT fails to extract the ORG entity “Foran” since it outputs a much lower similarity score. Our model successfully captures the semantics of “team competition” and it can be used to complement the



Strengthen	Complement	Weaken
S1: A beautiful Timber Frame bridge over a stream in Auburn (LOC), PA (LOC). Relational Triplet: (Auburn, contain, PA)	S2: Cross country: Foran (ORG)'s Mia Williams (PER), Kevin Preneta take firsts. Relational Triplet: (Mia Williams,member_of, Foran)	S3: Taylor Swift (PER) sets new AMAs (MISC) record, urges people to vote via @ReutersTV. Relational Triplet: (Taylor Swift, awarded, AMAs)
Similarity Score: 0.76 MDE Score - Strengthen: 0.954 Complement: 0.045 Weaken: 0.001	Similarity Score: 0.24 MDE Score - Strengthen: 0.000 Complement: 0.927 Weaken: 0.072	Similarity Score: 0.14 MDE Score - Strengthen: 0.000 Complement: 0.073 Weaken: 0.926
RpBERT: Auburn (LOC), PA (LOC) Ours: Auburn (LOC), PA (LOC)	RpBERT: Foran (PER), Mia Williams (PER) Ours: Foran (ORG), Mia Williams (PER)	HVPNeT: (/per/misc/present_in) Ours: (/per/misc/awarded)

Figure 5.5: The first line shows the three correlation categories, and the second row indicates representative samples with their ground-truth entity and relation types. The third line presents the comparison between our decomposed multimodal divergence estimation (MDE) score and conventional similarity score, and the bottom is the prediction results of our model and corresponding baselines.

missing semantics, which helps extract "Foran" as a name of organization and the relation "member_of" between the two entities. Another case is that when the image is irrelevant to textual contents in S3, HVPNeT gives the wrong prediction due to the misleading of the image. Our method can address this problem by generating a back-translation image of "Taylor Swift" and the "awarding scene", as shown in Figure 1.

5.6 Conclusion

We have revisited the misalignment issue in multimodal benchmarks. By borrowing the ideas from translation methods, we have implemented multimodal versions of back-translation and high-resource bridging, which provide a multi-view to the misalignment between modalities. The method has been validated in the experiments and outperforms 14 SOTA methods.

Chapter 6

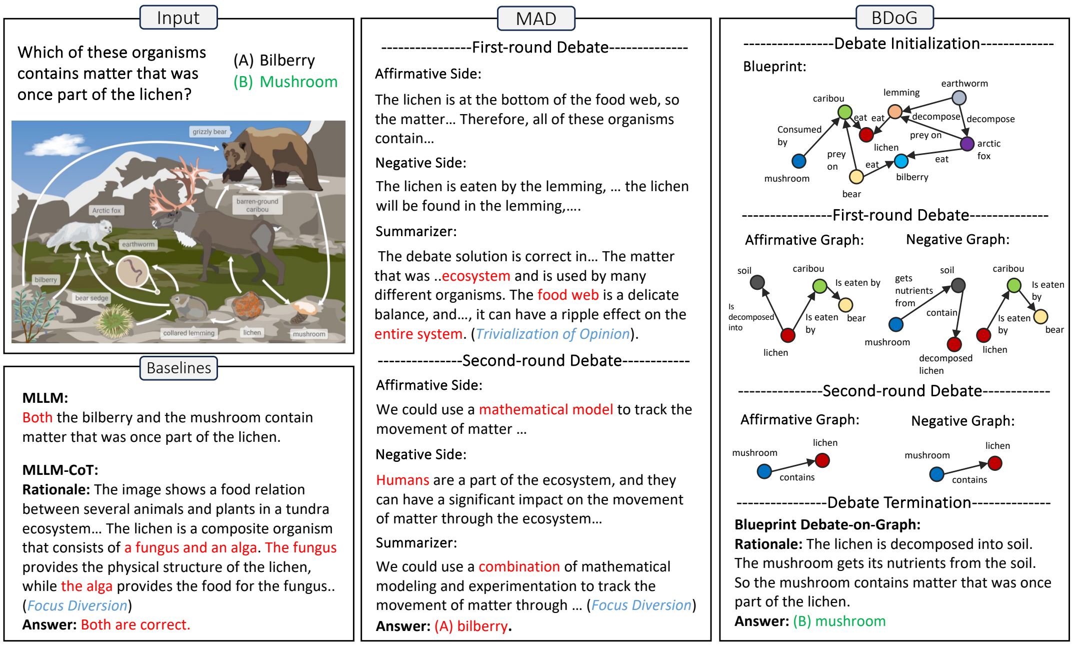
An Iterative Refined Blueprint Debate Paradigm for Knowledge Reasoning

6.1 Introduction

Multimodal reasoning depends on two key aspects: the creation of a unified representation of semantics from different modalities and the integration of these diverse semantics while ensuring logical consistency. While the advancement in large language models (LLMs) has made it possible to represent the semantics in natural languages [3, 98], the integration of diverse semantics remains a challenging issue, even in exclusive NLP tasks. One approach to tackle this challenge is multi-agent debate (MAD), where multiple LLMs act as agents, each contributing their own perspectives on the target topic and reaching a consensus through debates [64, 13]. This scheme could be adopted by incorporating a specific LLM for each modality as an agent.

While being relatively unexplored in the multimodal domain, MAD encounters numerous challenges in a broader context. It may suffer from the *trivialization* of opinions, resulting from the summarization step performed at the conclusion of each debating round. The objective of this step is to seek agreement among the participating agents regarding their opinions. Consequently, this process can lead to the debate's focus being directed towards a general concept, serving as an adaptation to accommodate the diverse range of semantics. One example can be observed in the reasoning of the Multimodal Large Language Model (MLLM) depicted in Figure 6.1, where the image modality presents a diverse range of semantics, including *bear sedge, earthworm, collared lemming*, and others. As a consequence, this can result in the context and summary being trivialized, shifting the emphasis from *lichen* to a more generalized concept of the tundra ecosystem, wherein both *bilberry* and *mushroom* exhibit a high degree of correlation. Similar issue exists when MAD is employed, where the summarizer concludes the diverse semantics into general words like *ecosystem* and *food web*, making the conclusion less specific. In addition, MAD may encounter the issue of *focus diversion*, which occurs when Chain-of-Thoughts (CoT) is utilized and new concepts introduced are highly correlated with a particular concepts (*e.g.*, *mathematical model* [19]), leading to an increased weighting of that concept within the context.

We argue that these challenges arise due to the inductive nature of existing debating schemes, wherein agent opinions are gathered from disparate concepts at word-level and consensus is achieved through bottom-up summarization. This approach may be effective in confined NLP tasks [36, 41], where the topic is often limited to a small number of concepts and the application of CoT remains constrained. However, in a multimodal scenario, certain modalities (*e.g.*, images) are information-rich and have a higher likelihood of introducing distracting concepts [73]. Consequently, it increases the semantic divergence within the context and the likelihood of trivialization. The semantic divergence increases further when the impacts of those concepts



Which of these organisms contains matter that was once part of the lichen?

(A) Bilberry
(B) Mushroom

MLLM:
Both the bilberry and the mushroom contain matter that was once part of the lichen.

MLLM-CoT:
Rationale: The image shows a food relation between several animals and plants in a tundra ecosystem… The lichen is a composite organism that consists of a fungus and an alga. The fungus provides the physical structure of the lichen, while the alga provides the food for the fungus.. (*Focus Diversion*)
Answer: Both are correct.

----------------First-round Debate--------------

Affirmative Side:

The lichen is at the bottom of the food web, so the matter… Therefore, all of these organisms contain…

Negative Side:

The lichen is eaten by the lemming, … the lichen will be found in the lemming,….

Summarizer:

The debate solution is correct in… The matter that was ..ecosystem and is used by many different organisms. The food web is a delicate balance, and…, it can have a ripple effect on the entire system. (*Trivialization of Opinion*).

---------------Second-round Debate------------

Affirmative Side:

We could use a mathematical model to track the movement of matter …

Negative Side:

Humans are a part of the ecosystem, and they can have a significant impact on the movement of matter through the ecosystem…

Summarizer:

We could use a combination of mathematical modeling and experimentation to track the movement of matter through … (*Focus Diversion*)

Answer: (A) bilberry.

----------------Debate Initialization-------------

Blueprint:

----------------First-round Debate--------------

Affirmative Graph: Negative Graph:

---------------Second-round Debate------------

Affirmative Graph: Negative Graph:

----------------Debate Termination--------------

Blueprint Debate-on-Graph:
Rationale: The lichen is decomposed into soil. The mushroom gets its nutrients from the soil. So the mushroom contains matter that was once part of the lichen.
Answer: (B) mushroom

Figure 6.1: Comparison results from ScienceQA dataset of direct answer from MLLM, Multimodal Chain-of-Thought (CoT), Multi-agent Debate (MAD) and our Blueprint Debate on Graph (BDoG). BDoG confines debates to a blueprint and stores evidence in graph branches, which mitigates word-level opinion trivialization and distractions caused by irrelevant concepts.

are amplified through CoT, particularly when the newly introduced concepts exhibit biases towards certain concepts, resulting in focus diversion.

To address these issues, we propose an deductive reasoning scheme called Blueprint Debate on Graph (BDoG, *pronounced bee-dog*). BDoG is inspired by the blueprint debate that has been employed in real-world debates, which distinguishes itself from other debates by its concentration on evaluating and refining a proposal (*e.g.*, blueprint) to address specific challenges or issues. BDoG begins by aggregating concepts from modalities and incorporating with their relationships into an initial graph. This graph serves as a blueprint that confines the scope of the discussion rather than having it open to irrelevant semantics as in existing schemes. More importantly, BDoG conducts the debate in a top-down manner by marking down conclusions on the graph.

This prevents trivialization as specific concepts are preserved rather than merged into general ones. This can be found from the example shown in Figure 6.1, where the scope is limited to the tundra ecosystem while specific concepts like *mushroom* and *lichen* are retained. Furthermore, the graph provide a compact and high-level guidance for the discussion process. The newly introduced concepts are incorporated into relevant branches instead of remaining as a word-level thoughts within the context. This reduces the likelihood of focus diversion since, in BDoG, the competition of semantics occurs at the branch level rather than the word level. This can be seen from Figure 6.1, where the most relevant branches related to the *soil* and *caribou* standout from the competition, eliminating the irrelevant semantics effectively. In addition to the advantages of scope-confined guidance and branch-level competition, BDoG also increases explainability, allowing for the tracking of discussion progress (Figure 6.1).

6.2 Problem Statement

Definition 1. ***(Blueprint-guided Reasoning):*** *Given question q, image i, context c, and a multi-agent system A, generate a reasoning path through iterative blueprint graph refinement to derive answer a.*

6.3 The Proposed Paradigm: BDoG

6.3.1 Preliminary

We begin by outlining existing approaches for tackling the multimodal reasoning problem. Figure 6.2 shows the specific distinction among them. Formally, given a question Q consisting of t tokens, our goal is to identify the correct answer A from a set of candidate answers. In the context of multimodal reasoning, the expected answer is intended to be inferred based on a visual context I and a textual clue C, in addition to the question itself.

Vanilla Prompting. Vanilla prompting approaches aim to predict an answer A by augmenting the input with illustrative examples D in addition to the question Q, visual context I, and textual clue C.

Multimodal CoT. As noted by Lu et al. [73], incorporating intermediate reasoning steps (rationales) can aid in predicting the correct answer, especially for complex multimodal reasoning tasks. To address this, we first generate a rationale $R = \{r_1, r_2, ..., r_k\}$ given the input. The generated rationale R is then concatenated with the original language input to update the language representation. This augmented language input is fed together with the original visual input I into the same model to infer the final answer.

DDCoT. The Duty-Distinct Chain of Thought framework proposes a novel approach for deconstructing questions into fundamental sub-questions, similar to breaking down reasoning into elementary steps. Contrary to prior work on conversational agents, Zheng et al. [132] employ the instruction to acquire the sub-question sequence $Q_1, Q_2, ..., Q_t$ in a single interaction. Within this framework, the final response A is obtained by aggregating the answers A_i to each sub-question Q_i and the generated CoT rationale R_i.

Self-Correction. Self-correction techniques [108] endeavor to iteratively enhance model predictions by leveraging feedback generated from the model itself. In particular, a *feedback* function $f : R \rightarrow R'$ is adopted to iteratively map model outputs to the refined responses.

MAD. MAD [64] presents a promising framework that fosters discursive exchange and cross-pollination of ideas between conversational models. Consider a debate comprising j rounds amongst a set of large language models acting as interlocutors, the *proponent* generates a rationale R'_p and response A_p revised in the light of rationales R_o presented by the *opponent* in prior turns.

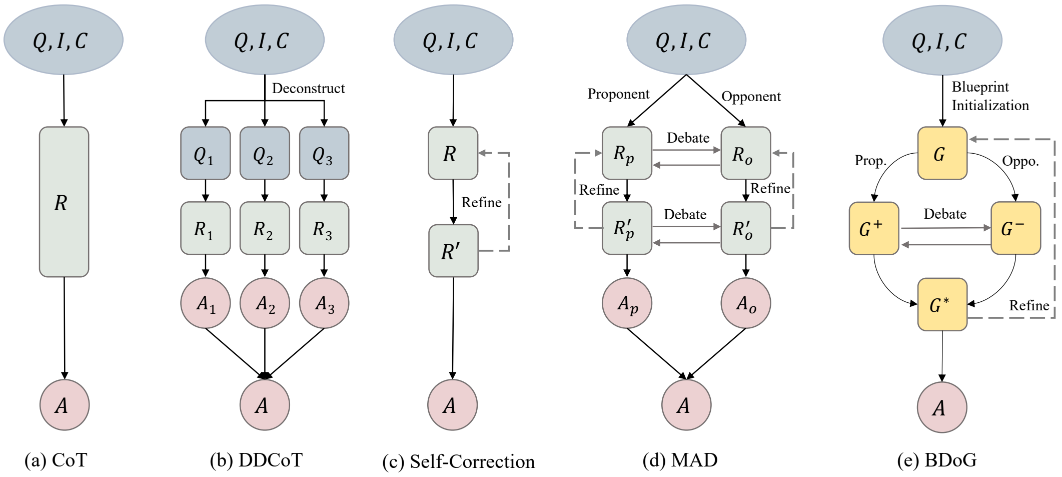


Figure 6.2: Comparison of CoT, Duty-Distinct CoT (DDCoT), Self-Correction, Multi-agent Debate (MAD) and Our proposed Blueprint Debate on Graph (BDoG). Q: input question, I: input image, C: context or hint, A: answer, R: rationale, G: blueprint.

6.3.2 Overall Architecture

In this section, we introduce Blueprint Debate-on-Graph (BDoG). As illustrated in Figure 6.2, BDoG takes a deductive approach instead of inducing answers from word-level thoughts. It utilizes graphs to structure the opinions and proposals provided by agents. This graph-level structuring of the debating context helps to minimize opinion trivialization and focus diversion. Furthermore, BDoG adopts a top-down approach which improves multimodal reasoning by iteratively refining an initial proposal, represented as a blueprint graph. This integrates opinions from diverse perspectives through the competition and cooperation among multiple agents.

The BDoG at the i^{th} round can be formulated as a quadruple

$$\mathcal{T}^i = (\mathcal{G}^i, \mathcal{S}, \mathcal{A}, \mathcal{F}) \tag{6.1}$$

where, given a multimodal source set $\mathcal{S} = \{Q, I, C\}$, the debating is conducted among a set of agents $\mathcal{A} = \{a_j\}, j \in \mathbb{Z}^+$, in which each agent uses operations from the set $\mathcal{F} = \{f_k\}, k \in \mathbb{Z}^+$ to propose opinions by refining the graph-of-thoughts $\mathcal{G}^i$. At the end of the i^{th} round, $\mathcal{G}^i$ is updated to $\mathcal{G}^{i+1}$ to initiate the next round.

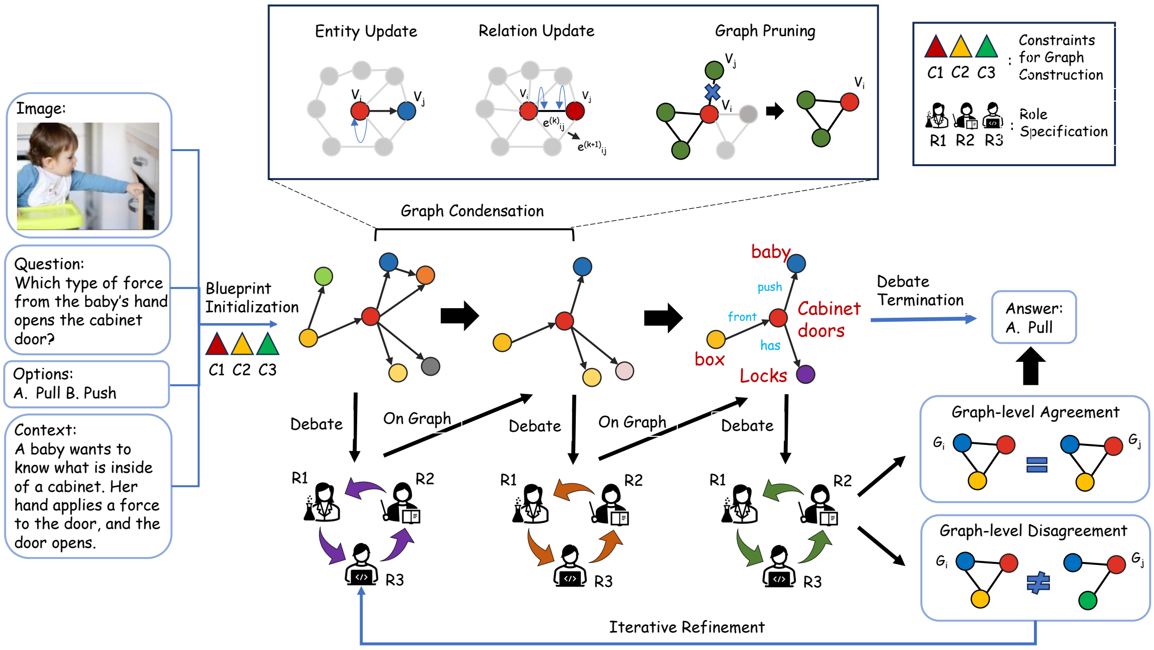


Image:

Question: Which type of force from the baby's hand opens the cabinet door?

Options: A. Pull B. Push

Context: A baby wants to know what is inside of a cabinet. Her hand applies a force to the door, and the door opens.

Answer: A. Pull

Figure 6.3: An overview of our Blueprint Debate-on-Graph (BDoG) framework. It iteratively refines the blueprint with a multi-agent debate paradigm.

6.3.3 Blueprint Initialization

To initiate the debating, we need to convert the multimodal sources into a blueprint graph. This conversion is achieved through the operation function $f_0 \in \mathcal{F} : \mathcal{S} \mapsto \mathcal{G}^0$. To implement f_0, we define two additional sub-functions f_t and f_v for extracting entities and relations from the textual sources (*i.e.*, Q and C) and visual source (*i.e.*, I), respectively. The implementation of f_0 is formulated as

$$\begin{aligned} f_0 :\ & \mathcal{S} \mapsto \mathcal{G}^0 \\ & \quad f_t(Q) \cup f_v(I) \cup f_t(C) \mapsto \langle \mathcal{V}^0, \mathcal{E}^0 \rangle \\ w.r.t\ & Size,\ Relevance \end{aligned} \tag{6.2}$$

where $\cup$ denotes the union of two sets of graphs. The 2 constraints are as follows: 1) **Size Constraint**: The size of $\mathcal{G}^0$ needs to be restricted within a specific range to prevent an excessive number of clues that could distract the inference or an insufficient number to answer the question effectively. 2) **Relevance Constraint**: We should merge the relationships extracted from I and C towards those of the question Q, ensuring all the knowledge encapsulated in $\mathcal{G}^0$ is relevant to the question. Extensive libraries are available for f_t and f_v, as they have been extensively researched (*e.g.*, named entity recognition [110], relation extraction [128] for f_t, image captioning [127], visual grounding [59] for f_v). However, the recent advancements in multimodal large language models (MLLM) have made it convenient to implement these sub-functions using in-context learning based prompts. For example, to extend the query I in the context, we can employ CoT to implement f_t as

> $f_t(Q)$: *Given the question {Q}, please provide the necessary steps to answer this question.*

where the { } denotes the placeholder in the prompt.

For $f_v(I)$, its implementation varies depending on LLMs used. For GPT-4, the image needs to be encoded in Base64 format. Gemini utilizes PIL for image encoding. InstructBLIP offers its EVA-G encoder to convert the image into an eigenvector. The f_0 can then be implemented as

> f_0: *Given the image {$f_v(I)$} and question {$f_t(Q)$}, generate a scene graph with evidence to answer the question. Please ensure adherence to following constrains: {Size}, {Relevance}.*

where two exemplar constraints are

> *Size : The graph must not be empty. Please restrict the maximum number of objects in the graph to 20.*

Relevance : The objects and relations within the graph should be pertinent to addressing the question.

It worth mentioning that although we provide some exemplar implementations of functions and constraints, the effectiveness of prompts can vary significantly depending on the MLLM used. The success of multimodal reasoning relies more on the development of guiding principles for prompting the models and constraints for regularizing the resulting graph. Therefore, in the rest of this section, our focus lies on discussing these guiding principles and constraints. Our prompt implementations will be provided in Appendix.

6.3.4 Agents and Roles

In the debate, we can treat each LLM as an agent that participates in the discussion by refining the blueprint graph $\mathcal{G}^0$. Just like in a real debate, each agent $a_j \in \mathcal{A}$ has a distinct role assigned. We define three roles as a set of $\mathcal{R} = \{Proponent, Opponent, Moderator\}$. These roles not only help structure the discussion but also promote critical thinking and ensure a comprehensive and in-depth exploration of the topic.

Proponent agents advocate and defend the current blueprint by refining current $\mathcal{G}^i$ into an affirmative evidence graph $\mathcal{G}^+$. A debating function is assigned for this purpose as

$$\begin{aligned} Proponent\ f^+ :\ & \mathcal{G}^i \times \mathcal{S} \mapsto \mathcal{G}^+ \\ & \langle \mathcal{V}^i, \mathcal{E}^i \rangle \cup f_t(Q) \cup f_v(I) \mapsto \langle \mathcal{V}^+, \mathcal{E}^+ \rangle \\ w.r.t\ & Size,\ Relevance,\ Compactness \end{aligned} \tag{6.3}$$

An exemplar implementation is

> f_+: *As* $\{personality\}$*, you are assigned as an* ***affirmative debater*** *and have been provided with an evidence graph* $\{\mathcal{G}^i\}$ *for answering the question* $\{f_t(Q)\}$ *related to the image* $\{f_v(I)\}$*. Try to enhance the graph by incorporating your insights towards an optimal solution. Please ensure adherence to following constrains:* $\{Size\}$*,* $\{Relevance\}$*,* $\{Compactness\}$*.*

Note that we have incorporated the conclusion from [24, 115] that the agent's understanding of the role can be improved by using the $\{personality\}$ for targeted personality injection. Furthermore, the personality can be tailored to be specific, such as "Ben, a high school student with an impressive academic record and respected by peers for your knowledge and logical thinking." The Proponent debate adheres to the Size and Relevance constraints defined in Eq. (6.2), and it also includes the **Compactness Constraint**: The refined graph should be as concise as possible, ensuring that the blueprint remains focused.

Opponent agents challenge and present arguments against the blueprint $\mathcal{G}^+$ by updating it into a negative evidence graph $\mathcal{G}^-$ as

$$\begin{aligned} Opponent\ f^- :\ & \mathcal{G}^+ \times \mathcal{S} \mapsto \mathcal{G}^- \\ & \langle \mathcal{V}^+, \mathcal{E}^+ \rangle \cup f_t(Q) \cup f_v(I) \mapsto \langle \mathcal{V}^-, \mathcal{E}^- \rangle \\ & w.r.t\ Size,\ Relevance,\ Compactness \end{aligned} \tag{6.4}$$

An exemplar implementation is

> f_+: *As* $\{personality\}$*, you are assigned as a* ***negative debater*** *and have been provided with an affirmative evidence graph* $\{\mathcal{G}^+\}$ *for answering the question* $\{f_t(Q)\}$ *regarding the image* $\{f_v(I)\}$*. Try to detect potential flaws and drawbacks of the graph and update it with your insights. Please ensure adherence to following constrains:* $\{Size\}$*,* $\{Relevance\}$*,* $\{Compactness\}$*.*

The utilization of the functions f_+ and f_- fosters an adversarial dynamic between the Proponent and Opponent, ensuring a diverse and comprehensive discussion.

To facilitate the debating, **Moderator** agents synthesize the arguments and opinions presented by both the proponent and opponent by merging the $\mathcal{G}^+$ and $\mathcal{G}^-$ into a conclusion $\mathcal{G}^*$ as

$$\begin{aligned} Moderator\ f_* :\ & \mathcal{G}^+ \cup \mathcal{G}^- \mapsto \mathcal{G}^* \\ & \langle \mathcal{V}^+, \mathcal{E}^+ \rangle \cup \langle \mathcal{V}^-, \mathcal{E}^- \rangle \mapsto \langle \mathcal{V}^*, \mathcal{E}^* \rangle \\ & w.r.t\ Size,\ Relevance,\ Compactness \end{aligned} \tag{6.5}$$

An exemplar implementation is

> f_*: *As {personality}, you are assigned as a **moderator** in a debate and have been provided with an affirmative evidence graph {$\mathcal{G}^+$} and a negative evidence graph {$\mathcal{G}^-$} to address the question {$f_t(Q)$} regarding the image {$f_v(I)$}. Try to consolidate the two graphs into a single graph towards the optimal solution, and provide a conclusive answer to the question.*

6.3.5 Debate Progress and Graph Condensation

Initialization and Role Assignment: Once the blueprint $\mathcal{G}^0$ has been initialized, the debate commences with the assignment of roles to agents in $\mathcal{A}$. Denote the assignment of a role $r \in \mathcal{R}$ to an agent a_j as $a_j := r$, to ensure a balanced debate, an equal number of agents are assigned as Proponents and Opponents, with only one agent assigned as the Moderator. The *Role Assignment Regulation* is

$$\begin{aligned} \big\|\{a_j | a_j := Proponent\}\big\| &= \big\|\{a_k | a_k := Opponent\}\big\|, \\ \big\|\{a_l | a_l := Moderator\}\big\| &= 1. \end{aligned}$$

Debating: After roles are assigned, the debate can be conducted iteratively between the Proponents and Opponents as illustrated in Figure 6.2. The initial blueprint $\mathcal{G}^0$ is then updated in subsequent debate rounds. In each round, the Moderator

Algorithm 2 BDoG

```
Input: Input $S$ = (question $Q$, image $I$ and context $C$), Multimodal LLM agents
  $A = (a_0, a_1, ..., a_n)$, Max debate round $R_{max}$.
  Initialize blueprint $G_0 \leftarrow$ Extract_Entity_Relation_Attribute $(a, S)$, proponent
  $a_p$, opponent $a_o$, and moderator $a_m$ with different personalities, $G \leftarrow G_0$.
  while $R \leq R_{max}$ do
    ▷ Affirmative Graph Generation
    $G_p \leftarrow$ Graph_Condensation $(a_p, G, S)$
    $G \leftarrow G_p$
    ▷ Negative Graph Generation
    $G_o \leftarrow$ Graph_Condensation $(a_o, G, S)$
    $G \leftarrow G_o$
    ▷ Debate Termination
    if $G_p = G_o$ or $R = R_{max}$ then
      $G \leftarrow [G_p, G_o]$
      Answer $(a_m, G, S)$
      break
    end if
  end while
```

summarizes the affirmative and negative graphs in a conclusion graph on the basis of which a tentative answer is also provided. If the debate is not concluded, the Moderator initiates the next round by assign $\mathcal{G}^-$ as the blueprint $\mathcal{G}^{i+1}$. Otherwise, the Moderator's answer is considered final and adopted.

Stopping Criteria: The condition to conclude the debate can be determined by assessing the modifications made to the evidence graph compared to the previous round as

$$d(\mathcal{G}^{i+1} - \mathcal{G}^i) \leq \epsilon \tag{6.6}$$

where d is a distance metric defined on the graphs. The rationale is that with each successful round of debate, the evidence becomes more concise, leading to the condensation of the evidence graph. Therefore, we can quantify the modification by tallying the number of entities (relations) that have been updated and pruned as

$$\begin{aligned} d(\mathcal{G}^{i+1} - \mathcal{G}^{i}) =& d(\langle \mathcal{V}^{i+1}, \mathcal{E}^{i+1} \rangle - \langle \mathcal{V}^{i}, \mathcal{E}^{i} \rangle) \\ =& d(\{\mathcal{V}^{i+1} \cap \mathcal{V}^{i}\}) + d(\{\mathcal{E}^{i+1} \cap \mathcal{E}^{i}\}) \\ &+ d(\{\mathcal{V}^{i} - \mathcal{V}^{i+1} \cap \mathcal{V}^{i}\}) + d(\{\mathcal{E}^{i} - \mathcal{E}^{i+1} \cap \mathcal{E}^{i}\}). \end{aligned} \tag{6.7}$$

6.4 Experimental Settings

6.4.1 Backbone Models

To evaluate its performance and generalizability, we have implemented Blueprint Debate-on-Graph (BDoG) using different prevalent multimodal large language models as backbones, including 1) **GeminiProVision** [98], an extensively parameterized model developed by Google, 2) **InstructBLIP** [20] and **LLaVA-v1.5** [68], which possesses more constrained dimensions and computational resources relative to alternative architectures, and 3) **GPT-4** [3] which is the fourth iteration of the GPT model developed by OpenAI.

6.4.2 Datasets and Metrics

In line with the general setup described in [132, 74], we perform our experiments using two extensively adopted multimodal question answering (QA) datasets. These datasets are widely recognized as standard benchmarks, specifically designed to evaluate the performance and effectiveness of models in addressing multimodal reasoning tasks. The two benchmarks are: 1) **ScienceQA-IMG** (SQA-IMG) [73] represents the first multimodal scientific question-answering corpus comprising 21,000 inquiries paired with multiple choices and accompanying images. As a *training-free* approach, we solely utilize the TEST and DEV partitions of ScienceQA-IMG following prior work [73] for comparative assessment. 2) **MMbench** [70] offers a more systematic and robust means for *zero-shot* reasoning evaluation compared to existing benchmarks such as VQAv2 [27] or COCO Captions [18]. We employ the official data split (MMBench-Dev) and code released by the originating authors. We report the accuracy metric through a heuristic matching procedure, following the same setting of the official benchmark [73]. Table 6.1 provides an overview of the size and diversity of datasets used in the work, including the number of instances, subjects, categories, and the average question length. These statistics can help in understanding the complexity and challenges posed by these datasets for multimodal reasoning-based QA systems.

Dataset	Instance	Subject	Category	Avg. Ques.
SQA-Test [73]	2017	3	65	9.3
SQA-Dev [73]	2097	3	66	9.6
MMBench-Dev [70]	4329	6	20	8.9

Table 6.1: The statistics of ScienceQA test and dev set and MMbench dev set. Avg. Ques. = average counts of tokens in questions.

6.4.3 Model Deployment

The specifics of model deployment and hyperparameter configurations for the InstructBLIP system are detailed in Table 6.2. Experimental evaluations are conducted leveraging the computational resources of two NVIDIA A100 GPUs. Consistent with prior work [20], we employ the Vicuna-13B as the large language model and the EVA-CLIP as the vision encoding module. It is noteworthy that InstructBLIP imposes constraints on the overall input length; consequently, we set the output limits for graph generation and candidate answer production to 128 and 50 tokens, respectively. For outputs exceeding 256 tokens in length, we apply truncation techniques.

Setting	Value
LLM	Vicuna-13B
Vision Encoder	EVA CLIP-G/14
Hardware Requirement	2x A100 (40GB)
Truncation Mode	Left
Number of Beams	5
Temperature	1.0
Top-p	0.9
Data Type	float32
Image Resolution	224x224
Maximum Input Length	256
Maximum Output Graph Length	128
Maximum Output Answer Length	50
Maximum Debate Round	4
Inference Time for SQA	4.7 s/sample
Inference Time for MMBench	5.4 s/sample

Table 6.2: Detailed model and experiment settings for InstructBLIP used in this work.

Furthermore, we report the inference time metrics for the ScienceQA (SQA) and MMBench datasets. Although the average question token count for MMBench is lower than SQA, the inference time required is higher. A plausible explanation for this discrepancy may be the inherently greater complexity of questions in the MMBench dataset, necessitating the generation of more intricate output graphs.

For the GeminiProVision and GPT-4V systems, we utilize their official APIs without employing any pre-processing or post-processing techniques.

Model	Size	SQA-IMG	MMBench
MiniGPT-4 [133]	7B	37.7	24.3
Qwen-VL [7]	7B	58.6 (67.1)	38.2
Qwen-VL-Chat [7]	7B	68.6 (68.2)	60.6
mPLUG-Owl2 [111]	8B	63.9	66.5
CogVLM-Chat [105]	17B	69.6	63.7
InstructBLIP [20]	13B	59.2 (63.1)	44.0
InstructBLIP+**BDoG**	13B	**63.5**	**55.8**
LLaVA-v1.5 [68]	13B	71.6	68.2
LLaVA-v1.5+**BDoG**	13B	**72.0**	**71.1**
GPT-3.5+CoT [107]	175B	67.4	-
GPT-3.5+DDCoT [132]	175B	72.5	-
GPT-4+CoT [107]	175B+	71.5	75.1
GPT-4+**BDoG**	175B+	**77.2**	**79.2**
GeminiProVision [98]	175B+	76.5	75.2
GeminiProVision+**BDoG**	175B+	**81.1**	**81.3**

Table 6.3: Overall zero-shot results on ScienceQA-IMG test set and MMBench dev set. Size = backbone model size. There are limited zero-shot results previously published on ScienceQA-IMG, so we reimplemented above models and report our findings. Where possible, we include results from the LLaVA paper for comparison (shown in parentheses). For MMBench, we refer to the scores listed on the official public leaderboard.

6.5 Results and Discussion

6.5.1 Performance Comparison to SOTA Methods

In contrast to the few-shot methodology, which exhibits susceptibility to the specific examples selected for training, we have opted for the zero-shot setting. This approach circumvents potential biases introduced by a limited sample size, ensuring a more

Model	**Method**	**ScienceQA-IMG-Dev**				**ScienceQA-IMG-Test**			
		NAT	SOC	LAN	Avg	NAT	SOC	LAN	Avg
MniGPT-4 [133]	Base	42.9	30.6	43.7	38.4	42.0	30.1	50.0	37.7
Qwen-VL [7]		52.1	59.8	58.3	55.0	55.7	62.0	77.3	58.7
Qwen-VL-Chat [7]		60.9	67.4	62.5	63.3	67.7	69.6	75.0	68.6
mPLUG-Owl2 [111]		60.6	68.0	45.8	62.8	62.5	66.2	61.4	63.9
CogVLM-Chat [105]		63.1	69.2	77.1	65.6	68.0	72.2	70.4	69.7
LLaVA-v1.5 [68]		66.1	74.9	72.9	69.4	70.1	74.2	81.8	71.9
InstructBLIP [20]	Base	53.7	57.3	47.9	54.8	58.1	61.0	61.4	59.2
	+ BDoG^{Debate}	59.7	55.6	**54.2**	58.1	**63.1**	58.2	72.7	61.4
	+ BDoG^{Graph}	58.1	61.3	52.1	59.0	60.6	62.6	68.2	61.5
	+ BDoG	**61.1**	**64.0**	52.1	**61.9**	61.1	**66.5**	**75.0**	**63.5**
GeminiProVision [98]	Base	68.9	81.6	75.0	73.7	72.9	81.5	88.6	76.5
	+ BDoG^{Debate}	73.3	81.1	77.1	76.2	75.3	82.8	**93.2**	78.5
	+ BDoG^{Graph}	69.8	84.8	**87.5**	75.6	74.7	86.8	88.6	79.6
	+ BDoG	**73.6**	**86.2**	85.4	**78.4**	**76.6**	**87.4**	93.2	**81.1**

Table 6.4: Ablation study on ScienceQA-IMG dev and test set. Question classes: NAT = natural science, SOC = social science, LAN = language science.

robust and generalizable model. We evaluate the proposed method by by comparing it against two sets of SOTA approaches as follows:

- **Open-Source Multimodal LLMs with Relatively Moderate Parameters** including MiniGPT-4 [133], Qwen-VL and Qwen-VL- Chat [7], CogVLM-Chat [105], mPLUG-Owl2 [111], LLaVA-v1.5 [68], and InstructBLIP [20], with parameter scales ranging from 7B to 17B.

- **Closed-Source Multimodal LLMs with Large-Scale Parameters**: GPT-3.5 [107], GPT-4V [3] and GeminiProVision [98]. Following the general standard, GPT-3.5 and GPT-4 have been incorporated with the CoT [107] or DDCoT [132] (built based on image captioning results). These models are known for their parameter scales above 175B and are considered to have the best performance.

The results are shown in Table 6.3. The integration of BDoG has resulted in a significant improvement across different backbones, as evidenced by the performance gains

Model	**Method**	**MMBench-Dev**						
		LR	AR	RR	FP-S	FP-C	CP	Avg
MniGPT-4 [133]	Base	7.5	31.3	4.3	30.3	9.0	35.6	24.3
Qwen-VL [7]		16.1	44.7	34.8	35.2	39.2	46.6	38.2
Qwen-VL-Chat [7]		32.2	59.8	43.5	66.2	48.3	79.4	60.6
mPLUG-Owl2 [111]		32.2	72.4	60.9	68.6	60.1	79.4	66.5
CogVLM-Chat [105]		29.7	65.8	60	66.9	58	76.7	63.7
LLaVA-v1.5 [68]		44.1	67.3	60.0	72.0	59.4	82.1	68.2
InstructBLIP [20]	Base	19.1	54.2	34.8	47.8	24.8	56.4	44.0
	+ BDoG^{Debate}	22.9	60.3	**52.2**	54.3	**28.0**	**68.9**	52.4
	+ BDoG^{Graph}	58.8	65.5	41.2	51.2	18.6	46.1	51.1
	+ BDoG	**63.3**	**71.9**	37.8	**56.3**	20.3	59.1	**55.8**
GeminiProVision [98]	Base	55.9	80.4	73.9	79.5	61.5	82.1	75.2
	+ BDoG^{Debate}	71.1	**85.1**	83.1	78.9	71.9	81.3	79.3
	+ BDoG^{Graph}	**75.0**	84.5	80.7	**81.4**	73.0	83.6	80.7
	+ BDoG	74.0	84.8	**83.4**	81.3	**73.7**	**84.4**	**81.3**

Table 6.5: Ablation study on MMBench dev set. Question classes: LR = Logical Reasoning; AR = Attribute Reasoning; RR = Relation Reasoning; FP-S = Fine-grained Perception (Single Instance); FP-C = Fine-grained Perception (Cross Instance); CP = Coarse Perception.

of 4.3% ∼ 5.7% on SQA-IMG and 6.1% ∼ 11.8% on MMBench. Notably, when combined with GeminiProVision, BDoG achieves SOTA performance on the ScienceQA-IMG test set and MMBench development set, achieving accuracies of 81.1% and 81.3%, respectively. Other observations that indicate BDoG's advantage over SOTA methods include:

BDoG helps reduce the performance gap between large and small models. It is commonly believed that models with larger parameter scales tend to perform better than smaller ones. This observation generally holds true, as shown in Table 6.3 for models without BDoG. However, the introduction of BDoG has led to a reduction in the performance gap between these two types of models. This can be seen in the improvement achieved by InstructBLIP, which has experienced a boost of 4.3% and achieves an accuracy of 63.5% on SQA-IMG, comparable to that of GPT-3.5. Similar results can be found in LLaVA-v1.5 with BDoG which gains the 71.1% accuracy in

MMBench, comparable to the GPT-4 model.

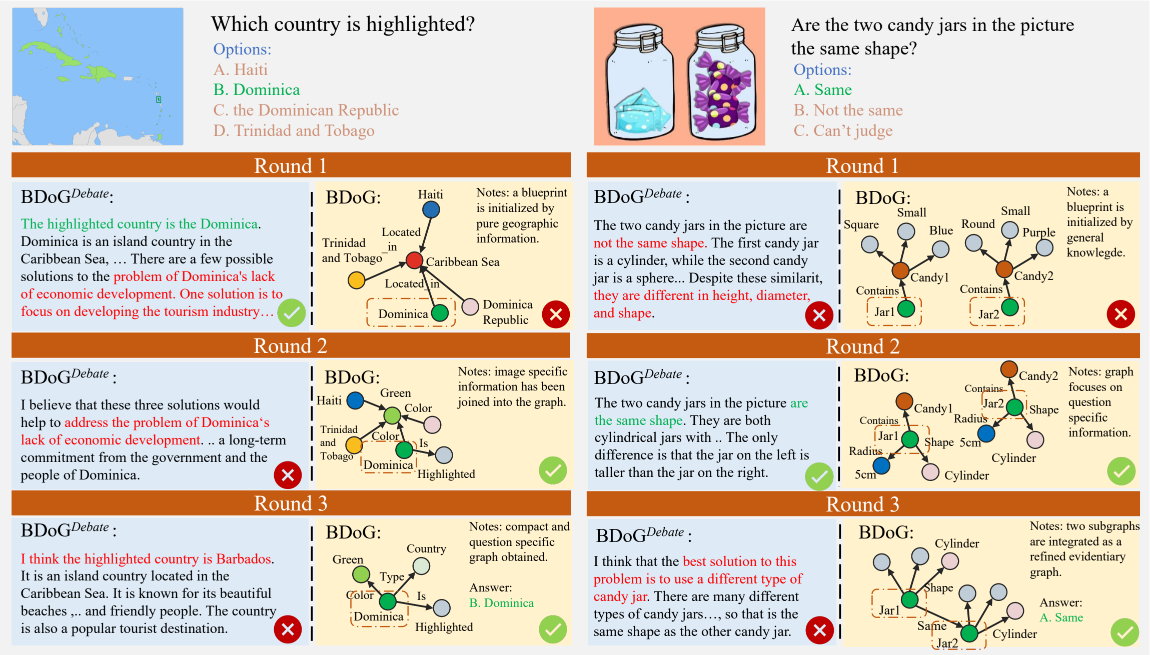
BDoG reinforces the multimodal reasoning. Form Table 6.3, we can also observe the advantage of direct multimodal reasoning (*e.g.*, open-source VL models, and GeminiProVision) over indirect multimodal reasoning (*e.g.*, GPT3.5+CoT and GPT3.5+DDCoT due to their nature of obtaining visual information through image captioning). Even the open-source VL models of the former group achieves comparable performance to those of the latter one, with much smaller parameter scales. With BDoG, which reinforces multimodal reasoning by graph regulation, the performance of direct multimodal reasoning of InstructBLIP and GeminiProVision have been improved by 6.1% and 11.8% on the MMBench dataset.

6.5.2 Ablation Study

In order to gain a comprehensive understanding of BDoG, we conduct an ablation study by decomposing BDoG into two variants:

- **BDoG**^{Debate}: we remove the graph regulation and constraints, resulting in a debate-only approach (*i.e.*, vanilla multi-agent debate) for investigating the specific contribution of the debating component of BDoG.
- **BDoG**^{Graph}: we remove the debating rounds, resulting in a graph-based reasoning method for investigating the specific contribution of the graph regulation component of BDoG.

Moreover, we analyze the performance of the two variants on the benchmarks by breaking it down into subcategories. This analysis allows us to investigate the preferences of these two variants for different types of questions. The results are presented in Table 6.5, where it can be observed that both variants demonstrate comparable performance across various benchmarks. This suggests that the debate and graph components of BDoG contribute to its effectiveness in a similar manner. Through the combination of these two components in BDoG, the performance has experienced further improvement compared to the individual variants. However, when considering specific categories, distinctions in the contributions of the debate and graph components become apparent.



Which country is highlighted?
Options:
A. Haiti
B. Dominica
C. the Dominican Republic
D. Trinidad and Tobago

Round 1

BDoG*Debate*:

The highlighted country is the Dominica. Dominica is an island country in the Caribbean Sea, … There are a few possible solutions to the problem of Dominica's lack of economic development. One solution is to focus on developing the tourism industry…

BDoG:

Notes: a blueprint is initialized by pure geographic information.

Round 2

BDoG*Debate* :

I believe that these three solutions would help to address the problem of Dominica‘s lack of economic development. .. a long-term commitment from the government and the people of Dominica.

BDoG:

Notes: image specific information has been joined into the graph.

Round 3

BDoG*Debate* :

I think the highlighted country is Barbados. It is an island country located in the Caribbean Sea. It is known for its beautiful beaches ,.. and friendly people. The country is also a popular tourist destination.

BDoG:

Notes: compact and question specific graph obtained.

Answer:
B. Dominica

Are the two candy jars in the picture the same shape?
Options:
A. Same
B. Not the same
C. Can't judge

Round 1

BDoG*Debate* :

The two candy jars in the picture are not the same shape. The first candy jar is a cylinder, while the second candy jar is a sphere... Despite these similarit, they are different in height, diameter, and shape.

BDoG:

Notes: a blueprint is initialized by general knowlegde.

Round 2

BDoG*Debate* :

The two candy jars in the picture are the same shape. They are both cylindrical jars with .. The only difference is that the jar on the left is taller than the jar on the right.

BDoG:

Notes: graph focuses on question specific information.

Round 3

BDoG*Debate* :

I think that the best solution to this problem is to use a different type of candy jar. There are many different types of candy jars…, so that is the same shape as the other candy jar.

BDoG:

Notes: two subgraphs are integrated as a refined evidentiary graph.

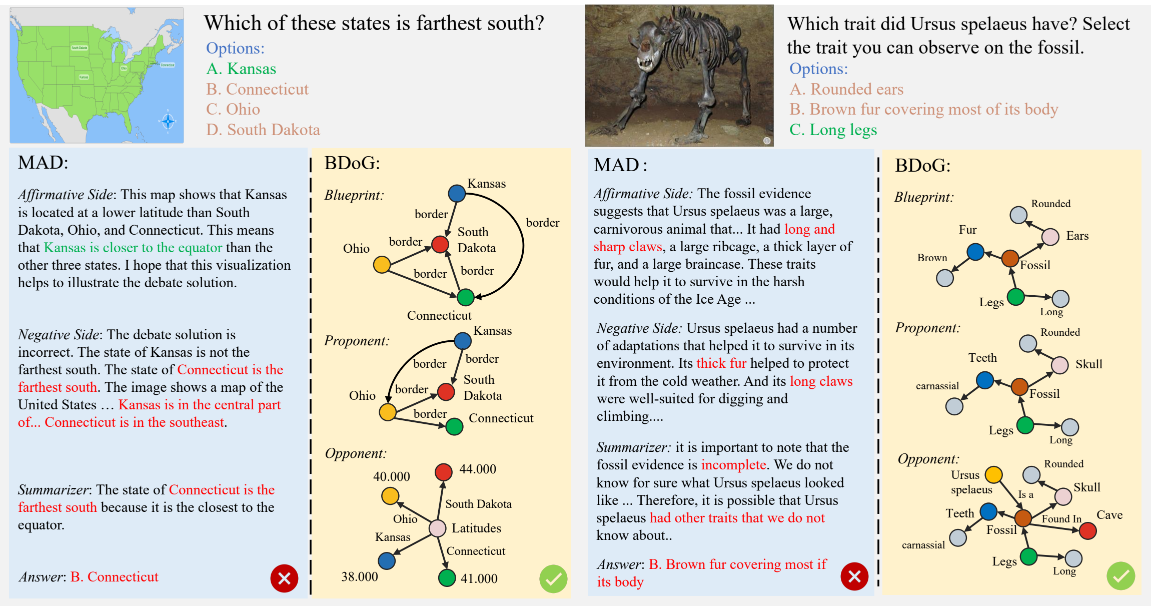
Answer:
A. Same

Figure 6.4: Case study of our proposed Blueprint Debate on Graph (BDoG) and vallina Multi-agent Debate (BDoG^{Debate}) on ScienceQA-IMG (left) and MMBench (right) datasets. Green color indicates the correct answer/rationale and Red means incorrect/irrelevant predictions.

Impact of the debate component: BDoG^{Debate} demonstrates consistent improvements across both benchmarks with a debate-only setting, which encourages LLM agents to collaboratively refine and correct prior responses. For science questions, **BDoG^{Debate}** facilitates the model's focus on specific errors, such as direction, size, and position, leading to improved performance in the natural science domain (boosting accuracy from 53.7 to 59.7 for InstructBLIP and 68.9 to 73.3 for GeminiProVision). However, the debate-only nature has limitations, including *trivialization* and *focus* *diversion* issues. Without the graph regulation, overall performance decreases from 55.8 to 52.4 for InstructBLIP, particularly when addressing questions that require attention to multi-hop logistic reasoning (LR) and specific attributes (AR).

Impact of the graph regulation: With a graph-regularized knowledge base for the discussion, $\mathbf{BDoG}^{Graph}$ also demonstrates consistent improvement of 2.3% $\sim$ 7.1% overs the base models on both benchmarks. Compared to the text-based and debate-only method $\mathbf{BDoG}^{Debate}$, it performs evidently better on the logistic reasoning and attributes reasoning questions by addressing the opinion trivialization and diversion with initialized blueprint. Although incorporating fact-related graph information proves beneficial in $\mathbf{BDoG}^{Graph}$, the absence of the iteratively refined debate procedure results in decreased performance due to the coarse and distorted extraction of blueprint information.

Impact of combining the debate and graph components: By combining the two components, BDoG achieves gains across nearly all categories. In the ScienceQA-IMG dataset, BDoG exhibits consistent and steady improvements, averaging around 5% compared to the baseline models. This suggests that BDoG is robust and generalizes well for science-related questions. Remarkably, BDoG significantly outperforms the baseline model (InstructBLIP) on the MMBench-Dev set, particularly in the areas of Logical Reasoning (LR) with a margin of 44.2%, Attribute Reasoning (AR) with a margin of 17.7%, and Relation Reasoning (RR) with a margin of 3%. BDoG enhances logical reasoning (LR) through a mechanism that refines the reasoning process iteratively, emphasizing the importance of multi-step reasoning rationales. The blueprint graph structure of BDoG, which explicitly models objects, attributes, and relations, contributes to improved reasoning abilities in Attribute Reasoning (AR) and Relation Reasoning (RR). The GeminiProVision model also exhibits comparable performance improvements, with BoG contributing to enhanced fine-grained perception across instances (FP-C), resulting in a gain of 12.2%. This improvement can be attributed to the connections established between various objects within the debate-on-graph framework.



Which of these states is farthest south?

Options:

A. Kansas

B. Connecticut

C. Ohio

D. South Dakota

MAD:

Affirmative Side: This map shows that Kansas is located at a lower latitude than South Dakota, Ohio, and Connecticut. This means that Kansas is closer to the equator than the other three states. I hope that this visualization helps to illustrate the debate solution.

Negative Side: The debate solution is incorrect. The state of Kansas is not the farthest south. The state of Connecticut is the farthest south. The image shows a map of the United States … Kansas is in the central part of... Connecticut is in the southeast.

Summarizer: The state of Connecticut is the farthest south because it is the closest to the equator.

Answer: B. Connecticut

BDoG:

Blueprint:

Proponent:

Opponent:

Which trait did Ursus spelaeus have? Select the trait you can observe on the fossil.

Options:

A. Rounded ears

B. Brown fur covering most of its body

C. Long legs

MAD :

Affirmative Side: The fossil evidence suggests that Ursus spelaeus was a large, carnivorous animal that... It had long and sharp claws, a large ribcage, a thick layer of fur, and a large braincase. These traits would help it to survive in the harsh conditions of the Ice Age ...

Negative Side: Ursus spelaeus had a number of adaptations that helped it to survive in its environment. Its thick fur helped to protect it from the cold weather. And its long claws were well-suited for digging and climbing....

Summarizer: it is important to note that the fossil evidence is incomplete. We do not know for sure what Ursus spelaeus looked like ... Therefore, it is possible that Ursus spelaeus had other traits that we do not know about..

Answer: B. Brown fur covering most if its body

BDoG:

Blueprint:

Proponent:

Opponent:

Figure 6.5: Intra-round case study comparing the proposed Blueprint Debate-on-Graph (BDoG) and vallina Multi-agent Debate (MAD) on ScienceQA-IMG (left) and MMBench (right) datasets. Green color indicates the correct answer/rationale and Red means incorrect/irrelevant predictions.

A case study for the iterative improvement on the blueprint: BDoG leverages the advantages of both structured evidence through graph regulation and iterative refinement through debating. This is evident in the consistent improvement observed on the blueprint graph, showcasing the combined benefits of these two components. Figure 6.4 provides running examples demonstrating the superior reasoning performance of our proposed BDoG framework compared to the **BDoG**^{Debate} method.

The left case draws from the ScienceQA dataset, testing geographic knowledge and map interpretation. While **BDoG**^{Debate} correctly answered *Dominica is highlighted*, it also generated irrelevant information about Dominica's economic development. This misguided the agents into off-topic discussion, concluding incorrectly with *Barbados*.

In contrast, BDoG concentrated on the question and options, iteratively refining the blueprint entities and relations to arrive at the right answer of Dominica.

The example on the right comes from the MMBench dataset requiring cross-instance perception. As the image contained both *candies* and *jars*, it posed a challenge. With **BDoG**^{Debate} relying on text alone, agreement was rarely reached as responses changed over debate rounds. However, BDoG first generated a blueprint defining image objects and attributes. This established the discussion scope. BDoG then pruned irrelevant *candy* information, focusing discussion on the specific object - *jars*. It output the final answer by comparing and connecting the two *jar* sub-graphs.

In summary, Figure 6.4 demonstrates that BDoG beats **BDoG**^{Debate} on both datasets through its blueprint-driven approach. This concentrates graph-based reasoning on salient topics and prunes irrelevant details to arrive at well-supported conclusions.

Figure 6.5 depicts the running examples within a debate round that compares BDoG with MAD on ScienceQA-IMG and MMBench datasets. Our proposed *Blueprint debate on graph (BDoG)* is a more effective way to present information than the vanilla *Multi-agent debate (MAD)*. It is more structured, more visual, and more interactive. This makes it easier to follow the flow of the debate, to identify the key points that are being made, and to explore the information in more detail.

The first case utilizes the ScienceQA dataset, evaluating geographic knowledge and map interpretation skills. The correct response (identifying the southernmost state) necessitates comparing the locations of various states. In the multi-agent debate scenario, despite the affirmative agent's accurate prediction of the relative position, the negative side overwhelmingly opposes it. Due to the widespread issue of hallucination in large language models (LLMs), detecting such misinformation demands significant effort. Instead of engaging in debate solely on the final answer, our model facilitates debate at the fact level. By strategically modifying blueprint nodes, BDoG demonstrates that the final answer should be derived by comparing the attributes (latitudes) of each candidate. BDoG offers the additional advantage of facilitating easier monitoring of changes, resulting in a more reliable system.

In the second case, drawn from the MMBench dataset, the task requires reasoning about attributes to answer the posed question. While the MAD method generates rationales incorporating extensive inherent knowledge and imagination, this leads to the erroneous inference of a live Ursus, distinct from the fossil depicted in the image. Notably, the question demands direct observation of the image, from which the correct answer – ”Long legs” – can be readily inferred. The extraneous information generated by MAD misguides the model towards irrelevant concepts. Conversely, our BDoG method commences by analyzing the image, the associated question, and the candidate options. This effectively restricts the scope of analysis to the attributes of the fossil. Subsequently, BDoG incorporates additional observed features and refines potentially inaccurate nodes, ultimately leading to the accurate prediction.

Round	*ScienceQA-IMG-Test*		*MMBench-Dev*	
	BDoG-S	BDoG-L	BDoG-S	BDoG-L
1	60.5	80.6	51.6	81.0
2	**63.5**	80.9	54.6	81.1
3	63.1	81.1	**55.8**	**81.3**
4	63.3	**81.4**	**55.8**	80.9

Table 6.6: Model performance with respect to the iteration round of debate. BDoG-S: InstructBLIP with BDoG, BDoG-L: GeminiProVision with BDoG.

6.5.3 Monitoring The Debating Progress

We evaluate the model’s performance against the termination criteria across multiple debate rounds based on the data in Table 6.6. Our analysis shows that for models with smaller parameters like InstructBLIP, moving from a single round to two rounds led

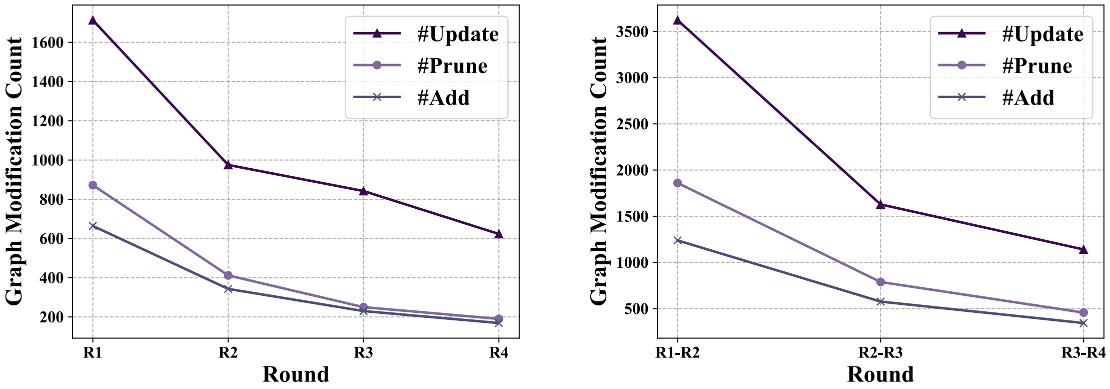


Figure 6.6: Statistics of intra-round (left) and inter-round (right) Blueprint condensation of BDoG with GeminiProVision for ScienceQA-IMG test set. #Update: number of updated attributes; #Prune: number of pruned entities/relations; #Add: number of newly-added entities/relations.

to significant gains in performance. This improvement is particularly notable when increasing the number of rounds from one to two. However, for larger models that may reach agreement more easily, the performance enhancement is relatively modest when amplifying the number of debate rounds. In general, we find the model's performance tend to converge within the second or third round. This can be attributed to the underlying reasoning typically being able to answer questions within 2-3 steps.

Additionally, Figure 6.6 illustrates the number of updated attributes, newly added or removed entities or relations between and within rounds. A strength of our proposed BDoG framework is its ability to quantify the debate process by inspecting graph changes. This demonstrates the effectiveness of dynamically adjusting the initial graph based on the discussion. The results in Figure 6.6 are also consistent with our hypothesis that disagreements and errors can be decreased as the debate progresses.

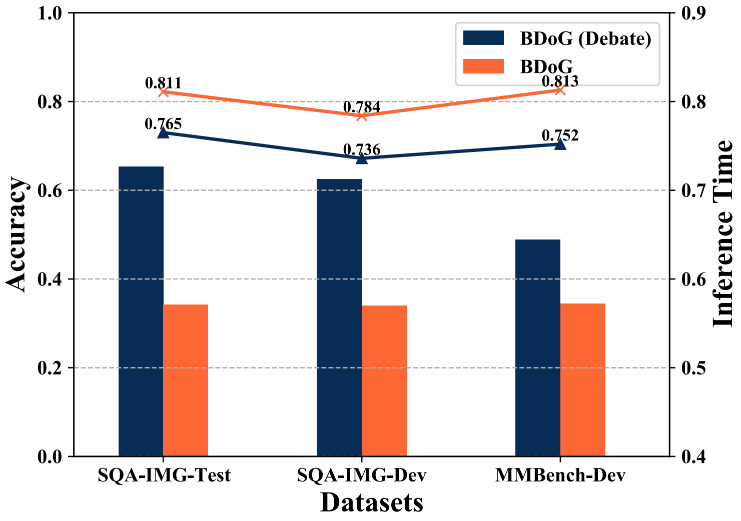


Figure 6.7: Effectiveness vs. efficiency results, comparing our proposed Blueprint Debate-on-Graph (BDoG) and vanilla Multi-agent Debate (BDoG (Debate)) on GeminiProVision. The bar chart indicates the inference time on three datasets and lines indicate the zero-shot performance (Accuracy).

6.5.4 Efficiency Analysis

We further compare the effectiveness versus efficiency of our BDoG framework against $\mathbf{BDoG}^{Debate}$, as shown in Figure 6.7. Maintaining concise content focuses on key aspects, the graph structure of BDoG demonstrates superior efficiency, requiring approximately 50% less inference time than $\mathbf{BDoG}^{Debate}$. By first generating a blueprint, BDoG defines the scope of the current state, thereby improving model efficiency by filtering irrelevant information. Concurrently, Figure 6.7 shows BDoG outperforms $\mathbf{BDoG}^{Debate}$ in effectiveness, achieving over 5 percentage higher accuracy than $\mathbf{BDoG}^{Debate}$ across three test sets. This enhanced effectiveness can be attributed to BDoG's concentrating on salient knowledge rather than generational textual content without guidance, as in $\mathbf{BDoG}^{Debate}$.

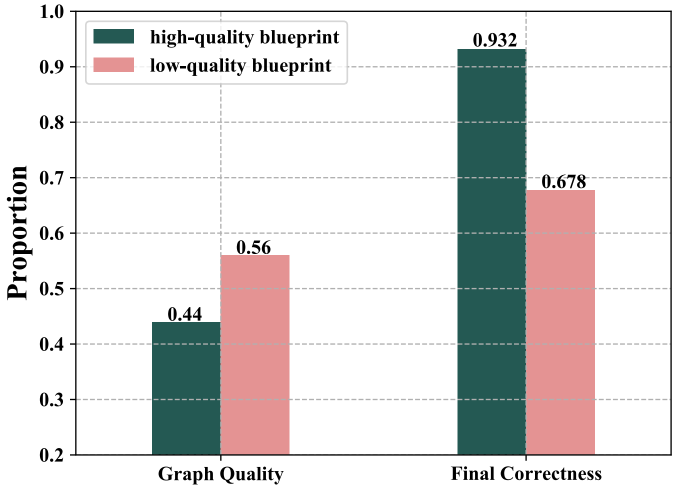


Figure 6.8: Human evaluation on the effect of blueprint quality for the GeminiPro-Vision model.

6.5.5 Effect of Blueprint Quality

We conduct a human evaluation to assess the impact of blueprint quality, as illustrated in Figure 6.8. A random sample of 200 predictions from the MMBench dataset is selected for evaluation. Due to the absence of a standardized metric for evaluating generated graph quality, three annotators are tasked with classifying each blueprint as either high or low quality. The results reveal that 56% of the initial blueprints were classified as low-quality. This finding aligns with expectations, given the complexity of the questions and the potential for the MLLM to generate coarse and imprecise direct answers.

To address the limitations of low-quality blueprints, we propose BDoG, a novel approach that iteratively refines the blueprint to enhance its conciseness and ultimately converge towards a correct answer. As demonstrated in the left panel of Figure 6.8, the final correctness rate is strongly correlated with blueprint quality. Notably, for questions with high-quality initial blueprints, the final correctness rate reaches

93.2%, highlighting the importance of a well-constrained initial graph. Furthermore, it is noteworthy that 67.8% of instances with low-quality blueprints ultimately result in correct predictions, demonstrating the effectiveness of BDoG's iterative refinement capabilities.

6.6 Conclusion

This work has presented a pioneering pilot study that introduces multi-agent debate into the realm of multimodal reasoning. We tackled two prominent challenges faced in this context: the issue of opinions being trivialized and focus diversion. By recognizing the limitations of existing debating schemes, we propose Blueprint Debate on Graphs (BDoG), which confines debates to a blueprint graph and stores evidence in graph branches, to address the challenges of word-level opinion trivialization and distraction caused by irrelevant concepts. Extensive experiments conducted in Science QA and MMBench validate the efficacy of BDoG, surpassing previous methods and establishing new state-of-the-art results.

Chapter 7

Conclusion and Future Work

In this thesis, we undertake a comprehensive investigation into four critical challenges in multimodal knowledge extraction and reasoning tasks: (1) achieving fine-grained alignment between visual and textual modalities, (2) addressing inconsistencies across different semantic levels, (3) mitigating biases and distortions in multimodal alignment data, and (4) enhancing dynamic refined knowledge graphs and knowledge-based reasoning. This chapter synthesizes our solutions to these challenges and suggests avenues for future research.

7.1 Conclusion

We introduce novel frameworks and methodologies designed to improve multimodal knowledge extraction and reasoning. Our contributions can be summarized as follows:

- MEGA (Multimodal Neural Network with Efficient Graph Alignment): By leveraging a sophisticated graph alignment method based on structural similarity and semantic agreement, MEGA effectively bridges the gap between visual and textual relations. This approach surpasses previous methods that relied on simple concatenation of graph representations, leading to superior alignment and knowledge extraction.

- RECK (REtrieval with Cross-modal Knowledge): RECK addresses the semantic gap between vision and language by incorporating high-level concepts from external knowledge graphs. This enriched semantic representation facilitates more accurate knowledge extraction, particularly in scenarios where traditional methods fall short due to simple concatenation of low-level features.

- TMR (Translation Motivated Multimodal Representation learning): TMR mitigates the negative impact of misalignment between images and text by employing Generative Back-translation and High-Resource Divergence Estimation. These components generate synthetic data to correct misalignments and quantify the degree of divergence, thereby enhancing the reliability of multimodal language understanding models.

- BDoG (Blueprint Debate on Graph): Inspired by real-world blueprint debates, BDoG introduces a deductive reasoning scheme that aggregates concepts from various modalities into an initial graph. This approach prevents the infiltration of irrelevant semantics and preserves specific concepts, effectively addressing the challenges of trivialization and focus diversion inherent in inductive reasoning schemes.

In conclusion, this thesis has contributed novel insights and practical solutions to the field of multimodal representation learning, moving us closer to the goal of universally effective multimodal AI systems capable of robust knowledge extraction and reasoning.

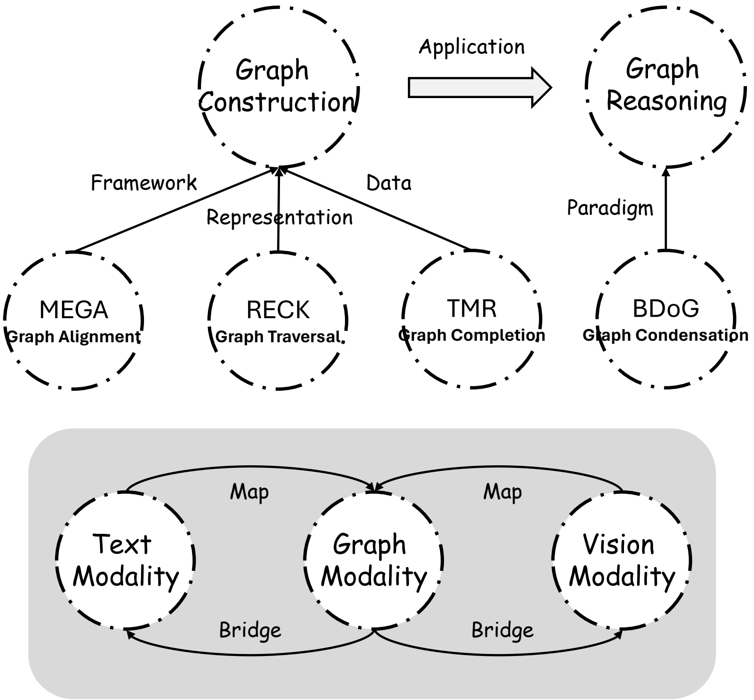


Figure 7.1: The interplay between our four works in this thesis.

7.2 Discussion

In this section, we discuss the relationship between our four multimodal representation methods within a unified graph learning framework, as illustrated in Figure 7.1.

The core thesis of our work posits that the graph modality serves as a bridge between vision and language by explicitly mapping different modalities into a unified multimodal knowledge graph. Each of our four approaches addresses distinct aspects of graph learning: alignment, traversal, completion, and condensation.

MEGA introduces an efficient method for semantic and structural graph alignment, which aligns graphs extracted from different modalities. This approach necessitates a fine-grained understanding of the provided images and sentences. Building on the aligned knowledge graph, **RECK** explores a graph traversal theory that retrieves knowledge paths linking visual objects and textual entities. These knowledge paths mimic human associative thinking, connecting concepts from abstract to specific.

Given that the extracted graph can be incomplete (due to biases and distortions), **TMR** addresses this issue by performing graph completion. This method supplements missing semantics by generating back-translation images as augmentation and estimating the divergence caused by such misalignments.

To overcome the limitations of static graphs, which hinder knowledge reasoning capabilities, **BDoG** constructs dynamic graphs. By iteratively refining the knowledge graphs (i.e., blueprints) using a multi-agent debate paradigm, **BDoG** significantly enhances multimodal reasoning abilities.

However, these methods have obvious limitation on efficiency issue. The complexity analysis is shown as follows:

- EGA: $O(|V_t||V_v|)$ for graph alignment

- RECK: $O(k|V|^2)$ for k-hop knowledge path extraction

- TMR: $O(mn)$ for m tokens and n image patches

- BDoG: $O(rd)$ for r debate rounds and d graph density

Although they are acceptable considering the performance improvement, we plan to optimize such efficiency issue in our future work.

Among these four methods, **MEGA**, **RECK**, and **TMR** primarily focus on the graph construction process. To demonstrate the effectiveness of the constructed multimodal knowledge graph, we employ the graph reasoning task as a real-world application. **BDoG** operates based on the given knowledge graph and reveals that graph-based multimodal reasoning holds potential for numerous downstream tasks requiring complex reasoning.

7.3 Future Work

The emergence of Large Language Models (LLMs) has revolutionized the field of Artificial Intelligence, particularly in Natural Language Processing (NLP). LLMs like GPT-4 and LLaMA exhibit remarkable capabilities in understanding and generating human-quality text, opening doors for a wide range of applications. One exciting avenue of exploration lies in leveraging LLMs for multimodal reasoning, where information from various modalities like text, images, and audio is integrated to achieve a more comprehensive understanding of the world. However, the complexity and heterogeneity of multimodal data pose significant challenges. Individual LLMs, while powerful, may not possess the specialized knowledge or reasoning abilities required to effectively process and integrate information across different modalities. This is where the concept of multi-agent collaboration becomes crucial.

In this thesis, we have explored the knowledge graph motivated multi-agent collaboration, including the multimodal knowledge graph construction (MEGA, RECK and TMR), and then we propose a novel paradigm called BDoG (blueprint debate on graph) that boost the multi-agent debate with a specified graph structure. Despite its potential, LLM-based multi-agent collaboration for multimodal reasoning presents several challenges that require further research: (1) Agent Communication and Coordination: Developing efficient and effective communication protocols for information exchange and action synchronization among agents. (2) Knowledge Representation and Sharing: Establishing a common knowledge representation framework that facilitates information sharing and understanding across different modalities and agents. (3) Learning and Adaptation: Enabling agents to learn from their interactions and experiences, improving their individual and collective performance over time. (4)

Evaluation Metrics: Defining appropriate metrics to evaluate the effectiveness of multi-agent collaboration in multimodal reasoning tasks.

Addressing these challenges will pave the way for the development of robust and versatile LLM-based multi-agent systems capable of tackling complex multimodal reasoning tasks, unlocking new possibilities in AI applications. In future work, we will go deeper into LLM-based multi-agent collaboration for multimodal reasoning. Specifically, we plan to conduct the research in the following directions:

(1) **Design and Develop a Collaborative Multi-agent Framework**: This framework will integrate multiple LLMs with diverse expertise in different modalities (e.g., vision, language, knowledge graphs) as collaborative agents, forming the foundation for multimodal reasoning.

(2) **Establish Effective Inter-agent Communication Protocols**: We will investigate and implement various communication protocols, such as structured messages, knowledge graphs, and natural language dialogues, to facilitate efficient information exchange and knowledge sharing among the LLM agents.

(3) **Develop Synergistic Reasoning Mechanisms**: This objective involves designing strategies for agents to combine their individual reasoning processes and insights to achieve a more comprehensive and accurate understanding of the problem.

(4) **Implement Conflict Resolution Techniques**: To address potential disagreements and inconsistencies among agents, we will explore and implement conflict resolution techniques such as negotiation strategies, voting mechanisms, and knowledge-based arbitration to reach a consensus and maintain consistency in reasoning.

(5) **Address Modality Gap and Data Limitation Challenges**: We will investigate techniques to bridge the modality gap, which arises due to the inherent differences in information representation across modalities. Additionally, we will address the challenge of limited training data by exploring techniques such as transfer

learning, multimodal pre-training, and data augmentation.

(6) **Prototype Development and Evaluation**: We will build a prototype system for multimodal question answering, showcasing the capabilities of the proposed framework. We will then rigorously evaluate its performance on benchmark datasets like ScienceQA-IMG and MMBench, comparing it to existing state-of-the-art approaches.

(7) **Explore Broader Applications**: We will investigate the applicability of the framework in other multimodal tasks such as image captioning, visual storytelling, and human-AI interaction, highlighting its potential for general-purpose multimodal reasoning and its versatility in addressing diverse challenges.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

[2] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545, 2014.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] Somak Aditya, Yezhou Yang, and Chitta Baral. Explicit reasoning over end-to-end neural architectures for visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[5] Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López-Monroy, and Thamar Solorio. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153, 2017.

[6] Peter Anderson, X. He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[8] Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, 2015.

[9] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*.

[10] Gregory Brown. An error analysis of relation extraction in social media documents. In *Proceedings of the ACL 2011 Student Session*, pages 64–68, 2011.

[11] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1989–1998, 2019.

[12] Marine Carpuat, Yogarshi Vyas, and Xing Niu. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79, 2017.

[13] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evalua-

tors through multi-agent debate. In *The Twelfth International Conference on Learning Representations*, 2023.

[14] Hong Chen, Yifei Huang, Hiroya Takamura, and Hideki Nakayama. Commonsense knowledge aware concept selection for diverse and informative visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 999–1008, 2021.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[16] Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. *arXiv preprint arXiv:2205.02357*, 2022.

[17] Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. Good visual guidance makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. *arXiv preprint arXiv:2205.03521*, 2022.

[18] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[19] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[20] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. In-

structblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pages 4171–4186, 2019.

[23] George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon, 2004.

[24] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

[25] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, 2018.

[26] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel P Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, 2011.

[27] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[28] Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor OK Li. Meta-learning for low-resource neural machine translation. In *2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 3622–3631. Association for Computational Linguistics, 2020.

[29] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(8):3549–3568, 2020.

[30] Wenya Guo, Ying Zhang, Xiangrui Cai, Lei Meng, Jufeng Yang, and Xiaojie Yuan. Ld-man: Layout-driven multimodal attention network for online news sentiment recognition. *IEEE Transactions on Multimedia*, 23:1785–1798, 2020.

[31] Zhijiang Guo, Yan Zhang, and Wei Lu. Attention guided graph convolutional networks for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 241–251, 2019.

[32] Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732, 2022.

[33] Xu Han, Tianyu Gao, Yuan Yao, Demin Ye, Zhiyuan Liu, and Maosong Sun. Opennre: An open and extensible toolkit for neural relation extraction. *arXiv preprint arXiv:1909.13078*, 2019.

[34] Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. OpenNRE: An open and extensible toolkit for neural relation extraction. In *Proceedings of EMNLP-IJCNLP: System Demonstrations*, pages 169–174, 2019.

[35] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, 2018.

[36] Jie He, Tao Wang, Deyi Xiong, and Qun Liu. The box is in the pen: Evaluating commonsense reasoning in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3662–3672, 2020.

[37] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[39] Mark Heimann, H. Shen, Tara Safavi, and Danai Koutra. Regal: Representation learning-based graph alignment. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018.

[40] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, 2010.

[41] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020.

[42] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[43] Sameera Horawalavithana, Sai Munikoti, Ian Stewart, and Henry Kvinge. Scitune: Aligning large language models with scientific multimodal instructions. *arXiv preprint arXiv:2307.01139*, 2023.

[44] Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[45] Yuting Hu, Liang Zheng, Yi Yang, and Yongfeng Huang. Twitter100k: A real-world dataset for weakly supervised cross-media retrieval. *IEEE Transactions on Multimedia*, 20(4):927–938, 2017.

[46] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*, 2023.

[47] Qingbao Huang, Jielong Wei, Yi Cai, Changmeng Zheng, Junying Chen, Ho-fung Leung, and Qing Li. Aligned dual channel graph convolutional network for visual question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7166–7176, 2020.

[48] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[49] Heng Ji and Ralph Grishman. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th annual meeting of the*

association for computational linguistics: Human language technologies, pages 1148–1158, 2011.

[50] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514, 2021.

[51] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[52] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[53] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[54] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, K. Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *HLT-NAACL*, 2016.

[55] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, 2016.

[56] Chenliang Li, Aixin Sun, Jianshu Weng, and Qi He. Tweet segmentation and its application to named entity recognition. *IEEE Transactions on knowledge and data engineering*, 27(2):558–570, 2014.

[57] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020.

[58] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[59] Kun Li, Jiaxiu Li, Dan Guo, Xun Yang, and Meng Wang. Transformer-based visual grounding with cross-modality interaction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6):1–19, 2023.

[60] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10313–10322, 2019.

[61] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[62] Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations*, 2023.

[63] Zejun Li, Zhihao Fan, Huaixiao Tou, Jingjing Chen, Zhongyu Wei, and Xuanjing Huang. Mvptr: Multi-level semantic alignment for vision-language pre-

training via multi-stage learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4395–4405, 2022.

[64] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

[65] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. Kagnet: Knowledge-aware graph networks for commonsense reasoning. 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)*, pages 2829–2839, 2019.

[66] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[67] Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji. A multi-lingual multi-task architecture for low-resource sequence labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 799–809, 2018.

[68] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[69] Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.

[70] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your

multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[71] Zuoguo Liu and Xiaorong Chen. Research on relation extraction of named entity on social media in smart cities. *Soft Computing*, 24(15):11135–11147, 2020.

[72] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999, 2018.

[73] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[74] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[75] Xuezhe Ma and E. Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *ArXiv*, abs/1603.01354, 2016.

[76] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, 2016.

[77] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

[78] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. *arXiv preprint arXiv:2311.17076*, 2023.

[79] Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. *arXiv preprint arXiv:2401.12863*, 2024.

[80] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. Multimodal named entity recognition for short social media posts. *arXiv preprint arXiv:1802.07862*, 2018.

[81] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307, 2017.

[82] Thien Huu Nguyen and Ralph Grishman. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, 2015.

[83] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[84] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL-HLT*, 2018.

[85] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[86] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in*

natural language processing, pages 1524–1534. Association for Computational Linguistics, 2011.

[87] Alan Ritter, Evan Wright, William Casey, and Tom Mitchell. Weakly supervised extraction of computer security events from twitter. In *Proceedings of the 24th International Conference on World Wide Web*, pages 896–905. International World Wide Web Conferences Steering Committee, 2015.

[88] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[89] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[90] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[91] Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, number FZJ-2022-00923. Jülich Supercomputing Center, 2021.

[92] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[93] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, 2019.

[94] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[95] Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations*, 2023.

[96] Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13860–13868, 2021.

[97] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3713–3722, 2020.

[98] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[99] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[100] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[101] Alakananda Vempala and Daniel Preoţiuc-Pietro. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 2830–2840, 2019.

[102] Keheng Wang, Feiyu Duan, Sirui Wang, Peiguang Li, Yunsen Xian, Chuantao Yin, Wenge Rong, and Zhang Xiong. Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering. *arXiv preprint arXiv:2308.13259*, 2023.

[103] Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. *arXiv preprint arXiv:2305.03453*, 2023.

[104] Linlin Wang, Zhu Cao, Gerard De Melo, and Zhiyuan Liu. Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307, 2016.

[105] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[106] Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21582–21592, 2023.

[107] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[108] Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. Generating sequences by learning to self-correct. In *The Eleventh International Conference on Learning Representations*, 2022.

[109] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton Van Den Hengel. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1367–1381, 2017.

[110] Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *Proceedings of the 28th ACM International conference on multimedia*, pages 1038–1046, 2020.

[111] Haiyang Xu, Qinghao Ye, Mingshi Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qiuchen Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Feiran Huang, and Jingren Zhou. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*, 2023.

[112] Pengcheng Yang, Fuli Luo, Peng Chen, Lei Li, Zhiyi Yin, Xiaodong He, and Xu Sun. Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling. In *IJCAI*, pages 5356–5362, 2019.

[113] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4145–4154, 2019.

[114] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019.

[115] Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuan-Jing Huang, and Xipeng Qiu. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15135–15153, 2023.

[116] Jianfei YU, Jing JIANG, Li YANG, and Rui XIA. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer.(2020). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352.

[117] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762, 2015.

[118] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In *Proceedings of COLING*

2014, the 25th international conference on computational linguistics: technical papers, pages 2335–2344, 2014.

[119] Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14347–14355, 2021.

[120] Hao Zhang, Gang Chen, Beng Chin Ooi, Kian-Lee Tan, and Meihui Zhang. In-memory big data management and processing: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 27(7):1920–1948, 2015.

[121] Jintian Zhang, Xin Xu, and Shumin Deng. Exploring collaboration mechanisms for llm agents: A social psychology view. *arXiv preprint arXiv:2310.02124*, 2023.

[122] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[123] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. Adaptive co-attention network for named entity recognition in tweets. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5674–5681, 2018.

[124] Yuhao Zhang, Peng Qi, and Christopher D Manning. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, 2018.

[125] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. Position-aware attention and supervised data improve slot filling. In

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 35–45, 2017.

[126] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[127] Wentian Zhao and Xinxiao Wu. Boosting entity-aware image captioning with multi-modal knowledge graph. *IEEE Transactions on Multimedia*, 2023.

[128] Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. Multimodal relation extraction with efficient graph alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5298–5306, 2021.

[129] Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. Mnre: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[130] Changmeng Zheng, Zhiwei Wu, Tao Wang, Cai Yi, and Qing Li. Object-aware multimodal named entity recognition in social media posts with adversarial learning. *IEEE Transactions on Multimedia*, 2020.

[131] Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. Progressive-hint prompting improves reasoning in large language models. *arXiv preprint arXiv:2304.09797*, 2023.

[132] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[133] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024.