

Variational Multi-Modal Hypergraph Attention Network for Multi-Modal Relation Extraction

Qian Li¹, Cheng Ji^{2,3*}, Shu Guo⁴, Kun Peng⁵, Qianren Mao³ and Shangguang Wang¹

¹School of Computer Science, Beijing University of Posts and Telecommunications, China

²SKLCCSE, School of Computer Science and Engineering, Beihang University, China

³Zhongguancun Laboratory, China

⁴National Computer Network Emergency Response Technical Team & Coordination Center, China

⁵Institute of Information Engineering, Chinese Academy of Sciences, China

li.qian@bupt.edu.cn, jicheng@act.buaa.edu.cn, guoshu@cert.org.cn, pengkun@ie.ac.cn, maoqr@zgclab.edu.cn, sgwang@bupt.edu.cn

Abstract

Multi-modal relation extraction (MMRE) is a challenging task that seeks to identify relationships between entities with textual and visual attributes. However, existing methods struggle to handle the complexities posed by multiple entity pairs within a single sentence that share similar contextual information (e.g., identical text and image content). These scenarios amplify the difficulty of distinguishing relationships and hinder accurate extraction. To address these limitations, we propose the variational multi-modal hypergraph attention network (VM-HAN), a novel and robust framework for MMRE. Unlike previous approaches, VM-HAN constructs a multi-modal hypergraph for each sentence-image pair, explicitly modeling high-order intra-/inter-modal correlations among different entity pairs in the same context. This design enables a more detailed and nuanced understanding of entity relationships by capturing intricate cross-modal interactions that are often overlooked. Additionally, we introduce the variational hypergraph attention network (V-HAN). This variational attention mechanism dynamically refines the hypergraph structure, enabling the model to effectively handle the inherent ambiguity and complexity of multi-modal data. Comprehensive experiments on benchmark MMRE datasets demonstrate that VM-HAN achieves state-of-the-art performance, significantly surpassing existing methods in both accuracy and efficiency.

1 Introduction

Relation Extraction (RE) is a fundamental task focusing on determining the relationships between pairs of entities within a given context [Xue *et al.*, 2022; Cao *et al.*, 2023]. Over recent years, there has been a growing interest in advancing this task through multi-modal relation extraction (MMRE), which

Evan Massey and Lori Blade presented with the AEO Meyer Harte Award










Entities	Objects
Evan Massey	  
Lori Blade	  
AEO Meyer	  

Figure 1: An example of the MMRE task. The task is to predict the relation of given entity pairs for the specific text and image, which contains multiple objects.

integrates textual information with visual data. By leveraging complementary insights from both modalities, MMRE addresses the limitations of text-only approaches, providing enriched semantic understanding and contextual grounding. MMRE plays a pivotal role in a variety of cross-modal applications. For instance, it unifies information across text and images to enable more comprehensive knowledge representation [Liang *et al.*, 2023; Ge *et al.*, 2021]. These advancements highlight MMRE’s potential to bridge the gap between textual and visual information, enabling innovative solutions for complex cross-modal challenges.

Previous studies [Chen *et al.*, 2022c; Liang *et al.*, 2022; Chen *et al.*, 2022d; He *et al.*, 2023] have made significant progress in advancing multi-modal relation extraction (MMRE) by effectively leveraging visual information to complement textual data. Visual content plays a crucial role in bridging semantic gaps, providing additional evidence to enhance the extraction of relationships. Despite their successes, these approaches primarily focus on aligning relations between objects and text, overlooking challenges posed by multiple entity pairs within a single sentence that share the same contextual information (e.g., identical text and image). This limitation makes it difficult to differentiate relationships between distinct entity pairs, leading to errors in scenarios where nuanced distinctions are critical. Addressing this gap remains an essential step toward advancing MMRE.

As illustrated in Figure 1, a single text-image pair can involve multiple entity pairs, each associated with distinct relationships. The most important objects for the entity *Evan Massey* and the *Lori Blade* are the objects framed in blue,

*Corresponding Author.

which are helpful to the task. Effectively associating entities with distinct sets of objects allows the model to extract meaningful semantic information that is critical for accurate relational classification. Additionally, a single entity may participate in multiple relationships across different pairs, introducing further diversity in the relations. For instance, *Evan Massey* appears in two distinct entity pairs, where some objects in the image contribute to multiple relationships. This overlap increases the complexity of learning specific representations, as the model must discern shared and unique contextual cues for each relation. This phenomenon underscores the importance of establishing accurate associations between entities and objects in the image for effective multi-modal relation classification. However, existing approaches often fall short of capturing correlations beyond simple pairwise interactions, especially when dealing with diverse entity relationships across modalities. Addressing this gap requires methods capable of modeling high-order correlations and diverse representations of entities and relations, a challenge that remains largely underexplored in the current landscape of MMRE research.

To address the challenges outlined above, we propose the variational multi-modal hypergraph attention Network (VM-HAN) for MMRE. Unlike existing methods that rely heavily on pre-defined features or rigid contextual structures, our approach dynamically learns a joint representation of multiple modalities by leveraging hypergraphs to capture complex, high-order correlations across different modalities. We model each sentence and its corresponding image (along with objects in the image) as a multi-modal hypergraph. This representation enables the model to go beyond pairwise interactions, capturing intricate, high-order relationships between textual and visual data. Our model autonomously learns the edges and weights of the hypergraph, optimizing its structure to uncover the latent relationships between entities and their associated visual and textual contexts. To further improve generalization and robustness, we adopt a variational approach, transforming node representations into Gaussian distributions. By modeling node features as distributions rather than fixed vectors, VM-HAN captures the underlying variability of relationships, resulting in more accurate predictions. The main contributions are summarized as follows:

- We technically design a novel MMRE framework to capture complex and high-order correlations among different modalities.
- We construct a multi-modal hypergraph capturing high-order correlations between different modalities. We also design Variational Hypergraph Attention Networks to handle the diversity of entities and relations.
- Experimental results on benchmark datasets demonstrate the effectiveness of our approach, outperforming state-of-the-art methods.

2 Related Work

2.1 Multi-Modal Relation Extraction (MMRE)

Multi-modal relation extraction (MMRE) has garnered significant attention in recent years [Zhang *et al.*, 2017; Wu *et al.*, 2020b; Zheng *et al.*, 2021a; Lu *et al.*, 2022; Cao *et al.*, 2021; Chen *et al.*, 2022a]. MMRE aims to identify textual relations between two entities in a sentence by incorporating visual content [Zheng *et al.*, 2021a; Zheng *et al.*, 2021b; Chen *et al.*, 2022c], which compensates for insufficient semantics and aids in relation extraction [Zheng *et al.*, 2021b; Zheng *et al.*, 2021a]. However, existing works ignore the multiple relations in different entity pairs in one sentence, which is caused by variations in entity pairs. In the task of multi-modal relation extraction, it is observed that images can provide valuable information. However, the potential of utilizing image information in a distinct manner for different entity pairs has not been fully explored. This paper aims to address this gap by capturing high-order correlations among entity pairs and the associated image objects and integrating intra-modal and inter-modal correlations that are truly useful for each entity pair.

et al., 2020b; Zheng *et al.*, 2021a; Lu *et al.*, 2022; Cao *et al.*, 2021; Chen *et al.*, 2022a]. MMRE aims to identify textual relations between two entities in a sentence by incorporating visual content [Zheng *et al.*, 2021a; Zheng *et al.*, 2021b; Chen *et al.*, 2022c], which compensates for insufficient semantics and aids in relation extraction [Zheng *et al.*, 2021b; Zheng *et al.*, 2021a]. However, existing works ignore the multiple relations in different entity pairs in one sentence, which is caused by variations in entity pairs. In the task of multi-modal relation extraction, it is observed that images can provide valuable information. However, the potential of utilizing image information in a distinct manner for different entity pairs has not been fully explored. This paper aims to address this gap by capturing high-order correlations among entity pairs and the associated image objects and integrating intra-modal and inter-modal correlations that are truly useful for each entity pair.

2.2 Hypergraph Network

Hypergraph networks have gained attention in various fields for modeling complex relationships among nodes [Gao *et al.*, 2023; Wu *et al.*, 2020a]. In the realm of multimodal learning, [Kim *et al.*, 2020] constructed a shared semantic space among different modalities by hypergraph and generated a joint representation by attentively integrating the modalities through a co-attention mechanism. MKHG [Zeng *et al.*, 2023] presents a degree-free hypergraph solution that ingeniously addresses the challenges posed by heterogeneous data sources and modalities. The effectiveness of hypergraph networks lies in their ability to capture rich semantic dependencies among entities, which is crucial in tasks such as multi-modal relation extraction. In this paper, we propose a novel approach that integrates hypergraph networks with multi-modal relation extraction to exploit high-order correlations among entity pairs and associated image objects.

3 Preliminaries

Multi-Modal Relation Extraction (MMRE). MMRE involves identifying relationships between entities by leveraging both textual and visual information. Formally, given a text $T = [w_1, w_2, \dots, w_i]$ associated with an image I , and an entity pair (h, t) , the goal is to predict the relationship between h and t from a predefined set of relations $\mathcal{R} = \{r_1, r_2, \dots, r_k, \text{none}\}$. An MMRE model takes the input (h, t, T, I) and outputs the most likely relation $r_i \in \mathcal{R}$. The output relationship inherently incorporates *visual cues* provided by the image I , enabling a richer contextual understanding. The model assigns a confidence score $p(r_i | h, t, T, I)$ for each possible relation r_i , determining which relation best describes the connection between the entities h and t in the given context.

Hypergraph. A hypergraph is a specialized form of a graph that differs from simple graphs by containing hyperedges. These hyperedges can connect two or more nodes and are often used to represent high-order correlations [Feng *et al.*, 2019]. A hypergraph is typically defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, consisting of a node set $\mathcal{V} = \{v_1, v_2, \dots, v_n\}$, a hyperedge set

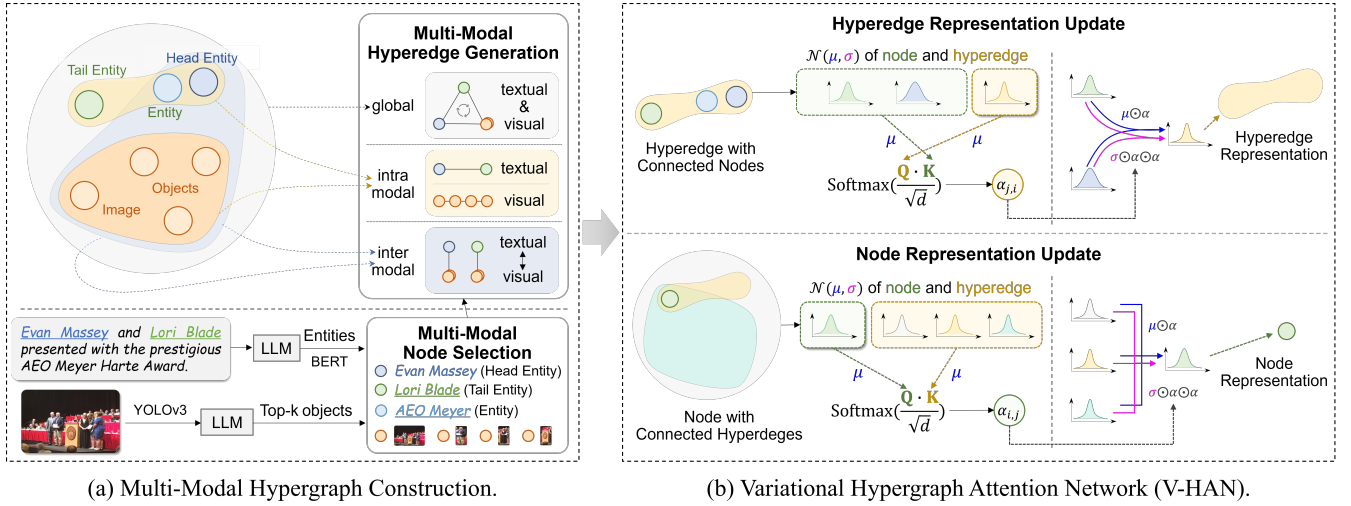


Figure 2: VM-HAN models text and corresponding images into a hypergraph for capturing high-order correlations and further learns entity pair representation under Gaussian distribution for robust nodes and hyperedges learning. The Multi-Modal Hypergraph Construction captures complex relationships among different modalities by creating global, intra-modal, and inter-modal hyperedges, and Variational Hypergraph Attention Networks (V-HAN) enhance representation diversity and robustness through variational modeling of nodes and hyperedges.

$\mathcal{E} = \{e_1, e_2, \dots, e_m\}$, and an optional diagonal weight matrix $\mathbf{P} \in \mathbb{R}^{m \times m}$ that represents weight of each hyperedge. The hypergraph \mathcal{G} can be represented by an incidence matrix $\mathbf{H} \in \{0, 1\}^{n \times m}$. In this way, each hyperedge e_j connects all associated nodes v_i , indicating correlations among them.

4 Framework

We propose the variational multi-modal hypergraph attention network (VM-HAN) framework, as illustrated in Figure 2.

4.1 Multi-Modal Hypergraph Construction

To effectively capture the high-order correlations across the modalities, we first construct a multi-modal hypergraph.

Multi-Modal Node Selection. The multi-modal hypergraph contains textual/visual nodes selected from the given text T and image I . Specifically, the entities h, t in text are initialized as textual nodes. We employ an LLM (LLama3) to extract the top m entities that are deemed useful for relation extraction. The prompt for this process is as follows:

Please identify and extract the top m entities and actions that are most relevant for relation extraction involving the given entity pair (h, t) .

Furthermore, we use an LLM (e.g., LLaVa) to process the corresponding image I and extract the top n entities deemed relevant, which is guided by the following prompt:

Please identify and extract the top n entities and actions from the image that are most relevant to the given entity pair (h, t) .

In addition, the visual nodes include the image itself and k most relevant objects to the given entity pair (h, t) . We utilize

another LLM (LLava), to select these k objects. The prompt for this process is as follows:

Please identify and select the top k objects in the image that are semantically most relevant to the given entity pair (h, t) .

Global Hyperedge. To capture global correlations among all modalities, we first construct the global hyperedges connecting all nodes. Formally, global hyperedge is as follows:

$$e_{\text{global}} = \{v_h, v_t, v_I, v_{o_1}, \dots, v_{o_k}\}. \quad (1)$$

Global hyperedge allows all nodes to propagate information to each other and get representations that contain global semantics (*i.e.*, both textual and visual information).

Intra-Modal Hyperedge. For relations within a single modality, we construct two intra-modal hyperedges for each hypergraph, one connecting the textual nodes (*i.e.*, head and tail entities) and the other connecting the visual nodes (*i.e.*, image and top- k objects). The intra-modal hyperedges are defined as follows:

$$e_{\text{textual}} = \{v_h, v_t\}, e_{\text{visual}} = \{v_I, v_{o_1}, \dots, v_{o_k}\}. \quad (2)$$

Intra-modal hyperedges focus on aggregating information within one single modality to obtain modal-specific semantics (*i.e.*, textual/visual correlations).

Inter-Modal Hyperedge. In addition, to capture the relations across different modalities, we construct two different inter-modal hyperedges, one connecting the head entity with the visual nodes, and the other connecting the tail entity with the visual nodes. The inter-modal hyperedges are defined as:

$$\begin{aligned} e_{\text{head-visual}} &= \{v_h, v_I, v_{o_1}, \dots, v_{o_k}\}, \\ e_{\text{tail-visual}} &= \{v_t, v_I, v_{o_1}, \dots, v_{o_k}\}. \end{aligned} \quad (3)$$

Inter-modal hyperedges mainly help to learn the associations between modalities (*i.e.*, text \leftrightarrow image). $e_{\text{head-visual}}$ denotes the cross-modal hyperedge connecting the head entity with the image and top three objects, and $e_{\text{tail-visual}}$ denotes the cross-modal hyperedge connecting the tail entity with the image and top three objects.

The advantages of our proposed multi-modal hypergraph construction module are twofold. First, by explicitly modeling the high-order correlations among different modalities, our module can capture more complex and fine-grained relations that might be overlooked by existing approaches. Second, by introducing hyperedges to connect multiple nodes simultaneously, our module can reduce the number of parameters needed to model the relationships among all nodes, which can reduce the risk of overfitting and improve the generalization performance of the model. The constructed multi-modal hypergraph contains three types of hyperedges, including global, intra-modal, and inter-modal hyperedges. To further study the strength of associations in the constructed multi-modal hypergraph, we next design an attention-based hypergraph encoder using variational modeling approaches.

4.2 Variational Hypergraph Attention Network

To further improve the performance of MMRE on the constructed multi-modal hypergraph, we propose a variational hypergraph attention network (V-HAN), utilizing a variational approach to learn representations under Gaussian distributions. The V-HAN model consists of two main components: variational hypergraph representation and variational hypergraph attention. The Variational Hypergraph Representation is responsible for encoding the nodes in the hypergraph, while the Variational Hypergraph Attention module is responsible for propagating the node information through hyperedges and updating the node representations.

Variational Hypergraph Representation. The variational hypergraph representation component serves as the foundation of V-HAN. It takes the textual and visual nodes as input and outputs representations modeled as Gaussian distributions. This representation captures both the core semantic features (mean) and the uncertainty (variance) of each node, providing a robust framework for modeling complex multi-modal relationships.

To obtain the node representations, we model the feature of each node $v_i \in \mathcal{V}$ using a Gaussian distribution and compute the mean and variance of node features for transferring specific representation into Gaussian representation as follows:

$$\mathbf{x}_{i,\mu}^{(0)} = \mathbf{W}_\mu \cdot \mathbf{x}_i, \quad \mathbf{x}_{i,\sigma}^{(0)} = \mathbf{W}_\sigma \cdot \mathbf{x}_i. \quad (4)$$

Specifically, we model each node v_i using a Gaussian distribution $N(\mathbf{x}_{i,\mu}, \text{diag}(\mathbf{x}_{i,\sigma}))$, where $\mathbf{x}_{i,\mu}^{(0)}$ and $\mathbf{x}_{i,\sigma}^{(0)}$ are the initial representations of the mean and variance of the node feature distribution, respectively. \mathbf{x}_i are the origin feature initialized in Section 4.1. $\mathbf{W}_\mu, \mathbf{W}_\sigma$ are learnable parameters. In this paper, we adopt the diagonal variance matrix, which is a common choice of previous works [Zhu *et al.*, 2019; Petrov, 2022].

The variational representations preserve the original feature information in the mean vector and estimate the uncer-

tainty of the original feature information in the variance vector. By modeling the node features as Gaussian distributions, this module provides a more informative and robust representation of each node, which helps for better multi-modal relation extraction performance.

Variational Hypergraph Attention. The variational hypergraph attention component updates node and hyperedge representations by leveraging attention mechanisms. This iterative process ensures that the model captures high-order correlations and dynamic associations between nodes and hyperedges. To adaptively learn the influences under the structure of multi-modal hypergraphs, we deploy a multi-head attention mechanism to compute the weight between hyperedge e_j and a node v_i connected with it as follows:

$$\alpha_{j,i}^{(l)} = \text{Softmax}_{v \in e_j} \left(\frac{(\mathbf{W}_e^{(l)} \mathbf{e}_{j,\mu}^{(l)}) \cdot (\mathbf{W}_x^{(l)} \mathbf{x}_{i,\mu}^{(l)})^T}{\sqrt{d}} \right), \quad (5)$$

where $\mathbf{x}_{i,\mu}$ is the mean vector of node v_i and $\mathbf{e}_{j,\mu}$ is the hyperedge representation updated below, and d is the dimension. The variational hypergraph attention mechanism updates the Gaussian distribution of hyperedge e_j by aggregating information from all nodes $v_i \in e_j$. Specifically, the mean and variance of hyperedge e_j in the $(l+1)$ -th layer of the updated distribution are computed using the following equations:

$$\begin{aligned} \mathbf{e}_{j,\mu}^{(l+1)} &= \sigma \left(\sum_{v_i \in e_j} \mathbf{x}_{i,\mu}^{(l)} \odot \alpha_{j,i}^{(l)} \cdot \mathbf{W}_{e,\mu}^{(l)} \right) + \mathbf{e}_{j,\mu}^{(l)}, \\ \mathbf{e}_{j,\sigma}^{(l+1)} &= \sigma \left(\sum_{v_i \in e_j} \mathbf{x}_{i,\sigma}^{(l)} \odot \alpha_{j,i}^{(l)} \odot \alpha_{j,i}^{(l)} \cdot \mathbf{W}_{e,\sigma}^{(l)} \right) + \mathbf{e}_{j,\sigma}^{(l)}, \end{aligned} \quad (6)$$

where $\mathbf{x}_{i,\mu}^{(l)}$ and $\mathbf{x}_{i,\sigma}^{(l)}$ are the mean and variance representation of node v_i in l -th layer. In the variational framework, especially one that deals with Gaussian distributions, each node (or hyperedge) is represented by a mean (μ) and a variance (σ^2). The mean represents the expected value of the feature, while the variance captures the uncertainty or spread of that feature's distribution. The double product in the updating of σ (variance) reflects the need to account for the uncertainty in the attention weights $\alpha_{j,i}^{(l)}$ themselves, in addition to the uncertainty in the node or hyperedge representations. This is particularly important in a setting where attention mechanisms are used, as the attention weights determine how much influence one node has over another. The squaring (or double product) of σ in this context ensures that the variance is scaled appropriately.

For node representation updating, similar to the hyperedges, we compute the weight between node v_i and a hyperedge e_j connected with it as follows:

$$\alpha_{i,j}^{(l)} = \text{Softmax}_{e \in v_i} \left(\frac{(\mathbf{W}_x^{(l)} \mathbf{x}_{i,\mu}^{(l)}) \cdot (\mathbf{W}_e^{(l)} \mathbf{e}_{j,\mu}^{(l)})^T}{\sqrt{d}} \right). \quad (7)$$

The updated Gaussian distribution of node v_i is computed by aggregating information from all hyperedges connected to it.

Specifically, the mean and variance of node v_i in the $(l+1)$ -th layer of the updated distribution are computed as follows:

$$\begin{aligned} \mathbf{x}_{i,\mu}^{(l+1)} &= \sigma \left(\sum_{e_j \in v_i} \mathbf{e}_{j,\mu}^{(l)} \odot \alpha_{i,j}^{(l)} \cdot \mathbf{W}_{x,\mu}^{(l)} \right) + \mathbf{x}_{i,\mu}^{(l)}, \\ \mathbf{x}_{i,\sigma}^{(l+1)} &= \sigma \left(\sum_{e_j \in v_i} \mathbf{e}_{j,\sigma}^{(l)} \odot \alpha_{i,j}^{(l)} \odot \alpha_{i,j}^{(l)} \cdot \mathbf{W}_{x,\sigma}^{(l)} \right) + \mathbf{x}_{i,\sigma}^{(l)}, \end{aligned} \quad (8)$$

where $\mathbf{e}_{j,\mu}^{(l)}$ and $\mathbf{e}_{j,\sigma}^{(l)}$ are the mean and variance representation of hyperedge e_j in l -th layer.

Through the iterative updating of attention weights and representations, V-HAN effectively captures high-order correlations across modalities, enabling the model to learn robust and informative representations for both nodes and hyperedges. By modeling these as Gaussian distributions, the module addresses challenges arising from ambiguity and variability in multi-modal data. Combined with the carefully designed joint optimization objectives presented in the next section, V-HAN significantly improves the performance of the proposed VM-HAN framework on MMRE tasks.

4.3 Joint Optimization Objectives

We design a joint loss function that evaluates the model's performance under the variational representation.

Relation Classification Constraint. The relation classification loss measures the error between the predicted relationship and the ground truth. This loss ensures that the model accurately classifies the relationship between entities in the multi-modal context, defined as follows:

$$\mathcal{L}_c = -\log p(r|\mathbf{x}_{h,\mu}, \mathbf{x}_{h,\sigma}, \mathbf{x}_{t,\mu}, \mathbf{x}_{t,\sigma}), \quad (9)$$

where $(\mathbf{x}_{h,\mu}, \mathbf{x}_{h,\sigma})$, $(\mathbf{x}_{t,\mu}, \mathbf{x}_{t,\sigma})$ are the representations of the two entities, and $p(r|\mathbf{x}_{h,\mu}, \mathbf{x}_{h,\sigma}, \mathbf{x}_{t,\mu}, \mathbf{x}_{t,\sigma})$ is derived by concatenating four vectors followed by a classification layer.

Reconstruction Constraint. The reconstruction loss \mathcal{L}_{rec} measures the difference between the predicted node representations and the actual node representations. It is defined as the mean squared error (MSE) between the predicted mean vectors $\mathbf{x}_{i,\mu}$ and the true mean vectors $\mathbf{x}_{i,h}$, and the MSE between the predicted variance vectors $\mathbf{x}_{i,\sigma}$ and the true variance vectors $\mathbf{x}_{i,\tau}$:

$$\mathcal{L}_{\text{rec}} = \frac{1}{n} \sum_{i=1}^n (|\mathbf{x}_{i,h} - \mathbf{x}_{i,\mu}|^2 + |\mathbf{x}_{i,\tau} - \mathbf{x}_{i,\sigma}|^2), \quad (10)$$

where n is the number of nodes.

Prior Constraint. The KL divergence loss \mathcal{L}_{KL} measures the difference between the predicted Gaussian distributions and a standard Gaussian distribution. It is defined as:

$$\mathcal{L}_{\text{KL}} = \frac{1}{n} \sum_{i=1}^n D_{\text{KL}}(\mathcal{N}(\mathbf{x}_{i,\mu}, \text{diag}(\mathbf{x}_{i,\sigma})) | \mathcal{N}(0, \mathbf{I})), \quad (11)$$

where D_{KL} is the Kullback-Leibler divergence and $\mathcal{N}(\mu, \Sigma)$ is the multivariate Gaussian distribution with mean μ and covariance matrix Σ . The KL loss plays a crucial role in regularizing the variational distribution, thereby enhancing the

generalization ability of the model. The overall loss function of VM-HAN is a weighted sum of the three loss terms:

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}}, \quad (12)$$

where λ_c , λ_{rec} , and λ_{KL} are hyperparameters that control the relative importance of each loss term.

5 Experiment

5.1 Experimental Setup

Dataset. To evaluate the performance of the proposed VM-HAN framework, we use two widely recognized datasets for Multi-Modal Relation Extraction (MMRE): MNRE and MORE. (1) The MNRE dataset [Zheng *et al.*, 2021b] is sourced from Twitter¹. (2) To broaden the scope of our investigation, we incorporate the MORE dataset [He *et al.*, 2023].

Comparison Methods. We compare VM-HAN against three categories of methods: text-based relation extraction (RE) models, BERT-based multi-modal RE (MMRE) models, and graph neural networks (GNNs) for the multi-modal relation extraction. (1) Text-based RE models include Glove+CNN [Zeng *et al.*, 2014], PCNN [Zeng *et al.*, 2015], Matching the Blanks (MTB) [Soares *et al.*, 2019]. (2) MMRE models include BERT+SG [Devlin *et al.*, 2019], BERT+SG+Att, VisualBERT [Li *et al.*, 2019], MEGA [Zheng *et al.*, 2021a], HVPNet [Chen *et al.*, 2022c], DGF-PT [Li *et al.*, 2023], MKGformer [Chen *et al.*, 2022b], MOREformer [He *et al.*, 2023], TMR [Zheng *et al.*, 2023], HVFormer [Liu *et al.*, 2024], CAMIM [Zhang *et al.*, 2024]. (3) Graph-based approaches include GCN [Kipf and Welling, 2016], GAT [Velickovic *et al.*, 2018], HGNN [Feng *et al.*, 2018]. More details can be found in Appendix B.2.

Implementation Details. For text-based initialization, the textual embeddings were initialized using the bert-base-uncased model from HuggingFace², with an embedding dimension of 768. Text inputs were either truncated or padded to a maximum sequence length of 128 tokens. For visual feature extraction, visual features were extracted using the VGG16 network³ and the YOLOv3 [Redmon and Farhadi, 2018], widely recognized for their performance in image feature extraction. The dimensionality of visual object features was set to 4096, and the number of objects per image was limited to three to maintain consistency and computational efficiency. The AdamW optimizer [Loshchilov and Hutter, 2019] was employed, with a learning rate of 2e-5 and a weight decay of 0.01. A dropout rate of 0.6 was applied to prevent overfitting, ensuring the model's robustness across diverse scenarios. The training process used a batch size of 16. To fine-tune hyperparameters effectively, a grid search approach was adopted, conducting five trials to identify the optimal configuration based on the validation set performance. Node embeddings in the multi-modal hypergraph were learned using a Graph Convolutional Network (GCN).

¹The direct link to the Twitter data stream provided at <https://archive.org/details/twitterstream>.

²<https://github.com/huggingface/transformers>

³<https://github.com/machrisaa/tensorflow-vgg>

Model Type	Model Name	MNRE dataset				MORE dataset			
		Acc. (%)	Prec. (%)	Recall (%)	F1 (%)	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)
Text-based RE	Glove+CNN [Zeng <i>et al.</i> , 2014]	70.32	57.81	46.25	51.39	60.23	27.87	31.65	29.64
	PCNN [Zeng <i>et al.</i> , 2015]	72.67	62.85	49.69	55.49	59.35	28.02	38.31	32.37
	MTB [Soares <i>et al.</i> , 2019]	72.73	64.46	57.81	60.96	60.19	29.40	40.37	34.02
Graph-based MMRE	GCN [Kipf and Welling, 2016]	73.64	63.70	67.41	65.50	78.28	59.57	59.10	59.33
	GAT [Velickovic <i>et al.</i> , 2018]	78.50	67.26	70.37	68.78	79.26	58.52	60.49	59.49
	HGNN [Feng <i>et al.</i> , 2018]	83.21	71.92	73.94	72.92	81.94	62.27	61.05	61.65
BERT-based MMRE	BERT+SG [Devlin <i>et al.</i> , 2019]	74.09	62.95	62.65	62.80	61.79	29.61	41.27	34.48
	BERT+SG+Att. [Devlin <i>et al.</i> , 2019]	74.59	60.97	66.56	63.64	63.74	31.10	39.28	34.71
	VisualBERT [Li <i>et al.</i> , 2019]	-	57.15	59.45	58.30	82.84	58.18	61.22	59.66
	MEGA [Zheng <i>et al.</i> , 2021a]	76.15	64.51	68.44	66.41	65.97	33.30	38.53	35.72
	HVPNet [Chen <i>et al.</i> , 2022c]	90.95	83.64	80.78	81.85	72.40	61.47	65.26	63.31
	MKGformer [Chen <i>et al.</i> , 2022b]	83.36	82.40	81.73	82.06	80.17	55.76	53.74	54.73
	MOREformer [He <i>et al.</i> , 2023]	82.67	82.19	82.35	82.27	83.50	62.18	63.34	62.75
	DGF-PT [Li <i>et al.</i> , 2023]	84.25	84.35	83.83	84.47	82.35	60.51	62.82	61.64
	TMR* [Zheng <i>et al.</i> , 2023]	87.35	84.48	83.66	84.07	83.19	62.57	64.70	63.62
	HVFormer [Liu <i>et al.</i> , 2024]	-	84.14	82.65	83.39	-	-	-	-
	CAMIM* [Zhang <i>et al.</i> , 2024]	89.42	84.27	84.90	84.58	83.42	63.32	65.15	64.22
	VM-HAN (Ours)	94.03 ($\uparrow 3.08$)	86.25 ($\uparrow 1.77$)	85.36 ($\uparrow 1.46$)	85.80 ($\uparrow 1.73$)	85.91 ($\uparrow 2.41$)	65.37 ($\uparrow 2.05$)	67.12 ($\uparrow 1.97$)	66.23 ($\uparrow 2.01$)

 Table 1: Main experiments. "—" means results are not available, and " \uparrow " means the increase compared to the second best baselines.

Variants	MNRE dataset					MORE dataset				
	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)	Δ Avg (%)	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)	Δ Avg (%)
VM-HAN (Ours)	94.03	86.25	85.36	85.80	-	85.91	65.37	67.12	66.23	-
w/o Multi-Modal Hypergraph	91.12	83.05	80.34	81.67	$\downarrow 3.82$	83.67	62.14	62.89	62.51	$\downarrow 3.36$
w/o Variational Representation	93.01	82.41	81.03	81.71	$\downarrow 3.32$	83.43	62.52	65.18	63.81	$\downarrow 2.43$
w/o V-HAN	89.36	81.50	81.59	81.54	$\downarrow 4.36$	81.52	60.69	61.40	61.05	$\downarrow 5.00$
w/o KL Loss	93.78	84.93	84.05	84.49	$\downarrow 1.05$	84.59	63.36	64.53	63.94	$\downarrow 2.06$
w/o Text LLM Enhancement	93.79	85.65	84.14	84.89	$\downarrow 0.74$	84.78	65.03	66.57	65.79	$\downarrow 0.62$
w/o Visual LLM Enhancement	93.90	85.69	85.13	85.41	$\downarrow 0.33$	85.54	65.07	66.42	65.73	$\downarrow 0.47$
w/o Global Hypergraph	93.68	85.43	82.32	83.85	$\downarrow 1.54$	84.52	63.76	65.38	64.56	$\downarrow 1.61$
w/o Intra-modal Hypergraph	93.61	84.77	84.72	84.74	$\downarrow 0.90$	84.90	63.78	65.81	64.79	$\downarrow 1.34$
w/o Inter-modal Hypergraph	93.25	84.18	83.44	83.81	$\downarrow 1.69$	84.55	63.12	65.33	64.21	$\downarrow 1.86$
repl. HGNN	89.01	81.45	80.43	80.94	$\downarrow 4.90$	81.76	61.64	63.75	62.69	$\downarrow 3.70$
repl. Variational GCN	89.58	81.10	81.25	81.18	$\downarrow 4.58$	82.59	60.30	62.51	61.39	$\downarrow 4.46$
repl. GCN	90.64	80.39	81.69	81.03	$\downarrow 4.42$	82.31	61.58	63.52	62.54	$\downarrow 3.67$
repl. GAT	90.62	81.84	81.31	81.57	$\downarrow 4.03$	82.93	61.35	63.16	62.24	$\downarrow 3.74$

Table 2: Ablation study. "w/o" means removing the corresponding module and "repl." means replacing the corresponding module.

5.2 Main Results

To evaluate the effectiveness of our proposed model, we compared it against SOTA baselines, some results were sourced from the respective original publications [Zheng *et al.*, 2021a; Chen *et al.*, 2022c; Li *et al.*, 2023; He *et al.*, 2023] for consistent comparisons. From Table 1, we can observe that: 1) Our model outperformed all baseline methods, confirming its ability to effectively integrate multi-modal knowledge for improved performance. 2) Across the four evaluation metrics, our model consistently demonstrated superior performance. Notably, it achieved a minimum improvement of 1.08% in F1-score and 2.93% in Accuracy, underscoring its robustness and efficiency. 3) When compared to text-based RE methods, our model demonstrated clear advantages, highlighting its capacity to utilize visual information to enhance relational understanding. 4) Our model surpassed BERT-based MMRE models, emphasizing its ability to capture intricate structural features through hypergraph learning techniques. A particularly noteworthy observation is the significant performance gains on the MORE dataset, where our model achieved at least a 2.93-point improvement across the evaluation metrics.

This improvement can be attributed to the dataset’s richness in visual objects, which enables the model to capture more fine-grained cross-modal relationships.

5.3 Ablation Study

From Table 2, we can observe that: 1) All of the core modules of VM-HAN demonstrate significant improvements in performance. 2) The multimodal hypergraph emerged as a particularly influential component, with its removal causing the most substantial performance drop among the tested variants. This can be attributed to its ability to model high-order relationships, thereby providing richer contextual clues for relation extraction. 3) Replacing the variational representations with fixed, specific representations led to noticeable performance declines. This suggests that the variational framework enables the model to better capture the underlying distributions of relationships. 4) Removing any one type of hypergraph (global, intra-modal, or inter-modal) resulted in a clear decrease in performance. This demonstrates that each hypergraph type captures distinct high-order relationships, all of which are essential for comprehensive multimodal reasoning.

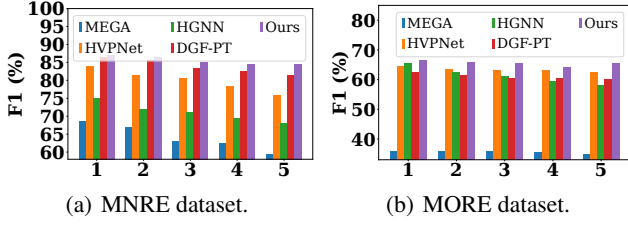


Figure 3: Impact of differences in sample number. It means the performance when an entity belongs to one or multiple entity types.

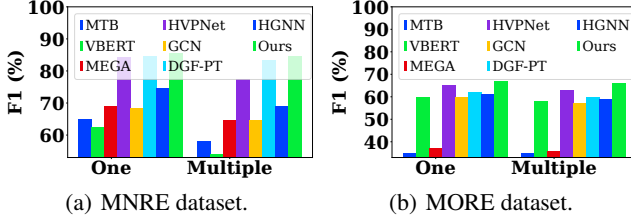


Figure 4: Impact of relation numbers for each sentence. It means the performance when a sentence has one or multiple relations.

5.4 Discussions for V-HAN

Entities Appearing in Multiple Categories. In this scenario, an entity is associated with multiple categories across different occurrences. We compared the performance of our proposed method against a baseline on datasets with varying numbers of entity repetitions, as shown in Figure 3. The results demonstrate that our method outperformed the baseline in both single and multiple repetition scenarios. This indicates that V-HAN effectively models the diversity of multi-semantic entities, capturing nuanced semantic representations that enhance the accuracy of relation extraction.

Entities with Different Meanings Across Pairs. Another challenging scenario involves entities that take on different meanings when paired with various counterparts in the same sentence, leading to different relation types. We evaluated the performance of our method and baselines on sentences containing different numbers of entity pairs, as illustrated in Figure 4. Our model consistently achieved state-of-the-art performance in both single-pair and multi-pair scenarios. This suggests that the hypergraph learning mechanism in V-HAN successfully leverages inter-modal associations, effectively distinguishing the differences across entity pairs.

5.5 Effect of Visual Information

To evaluate the contribution of visual information, we conducted an experiment comparing its performance with and without image features. The results are shown in Figure 5. Specifically, we trained two versions of the VM-HAN model: one utilizing both textual and visual data, and the other relying solely on textual data. From the figure, we can observe that the results clearly indicate that the VM-HAN model incorporating visual features outperforms the text-only version across all evaluation metrics. By integrating visual features,

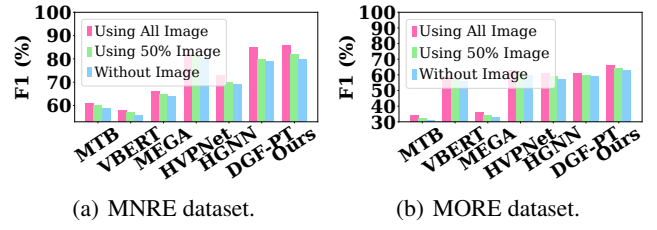


Figure 5: Different proportions of visual information.

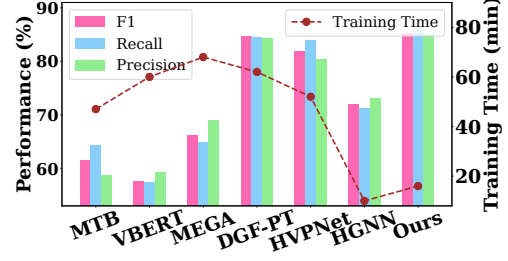


Figure 6: Performance and efficiency.

VM-HAN effectively captures complex, high-order correlations between modalities, which allows for more precise relation extraction in a multi-modal context.

5.6 Efficiency

To evaluate the efficiency of VM-HAN, we conducted a comparative analysis against several baseline models, with the results presented in Figure 6. The results indicate that our method achieves significant improvements in training efficiency while maintaining state-of-the-art accuracy. In comparison to BERT-based models, which are widely adopted but computationally demanding, our approach demonstrates substantially reduced training time. Unlike traditional graph neural networks, our hypergraph-based design captures high-order features both within and across modalities, streamlining the learning process and enabling faster convergence. The dual advantage of accuracy and efficiency positions ours as a practical and effective solution for real-world applications.

Case Study. We also provide a case study to showcase the ability of MV-HAN to identify and utilize relevant visual information effectively in Appendix C.

6 Conclusion

In this work, we introduced the variational multi-modal hypergraph attention network (VM-HAN), a novel framework that effectively addresses the challenges of multimodal relation extraction (MMRE). By incorporating hypergraph structures and variational modeling, our approach captures complex, high-order correlations across modalities, enabling it to model intricate associations between entities and their relationships. The variational modeling technique employed by our variational hypergraph attention Networks (V-HAN) proves particularly effective in handling polysemous entities. Extensive experimental evaluations demonstrate that our proposed VM-HAN achieves state-of-the-art performance.

Acknowledgments

We thank the anonymous reviewers for their insightful comments and suggestions. The corresponding author is Cheng Ji. The authors of this paper were supported by the NSFC through grant No.62402054, No.62425203 and No.62032003, and the 76th batch of general grants from China Postdoctoral Science Foundation through grant 2024M760279, and supported by the Postdoctoral Fellowship Program and China Postdoctoral Science Foundation under Grant Number.

References

- [Cao *et al.*, 2021] Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, and Wei Bi. Uncertainty-aware self-training for semi-supervised event temporal relation extraction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21*, page 2900–2904, New York, NY, USA, 2021. Association for Computing Machinery.
- [Cao *et al.*, 2023] Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. Zero-shot cross-lingual event argument extraction with language-oriented prefix-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12589–12597, 2023.
- [Chen *et al.*, 2022a] 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. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 904–915, 2022.
- [Chen *et al.*, 2022b] 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. In *SIGIR*, pages 904–915, 2022.
- [Chen *et al.*, 2022c] Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. Good visual guidance make A better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *NAACL*, pages 1607–1618, 2022.
- [Chen *et al.*, 2022d] Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web conference 2022*, pages 2778–2788, 2022.
- [Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.
- [Feng *et al.*, 2018] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. *CoRR*, abs/1809.09401, 2018.
- [Feng *et al.*, 2019] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *AAAI*, volume 33, pages 3558–3565, 2019.
- [Feng *et al.*, 2024] Yifan Feng, Shuyi Ji, Yu-Shen Liu, Shaoyi Du, Qionghai Dai, and Yue Gao. Hypergraph-based multi-modal representation for open-set 3d object retrieval. *TPAMI*, 46(4):2206–2223, 2024.
- [Gao *et al.*, 2023] Yue Gao, Yifan Feng, Shuyi Ji, and Rongrong Ji. Hggn⁺: General hypergraph neural networks. *TPAMI*, 45(3):3181–3199, 2023.
- [Ge *et al.*, 2021] Xuri Ge, Fuhai Chen, Joemon M. Jose, Zhilong Ji, Zhongqin Wu, and Xiao Liu. Structured multi-modal feature embedding and alignment for image-sentence retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia, MM '21*, page 5185–5193, New York, NY, USA, 2021. Association for Computing Machinery.
- [He *et al.*, 2023] Liang He, Hongke Wang, Yongchang Cao, Zhen Wu, Jianbing Zhang, and Xinyu Dai. MORE: A multimodal object-entity relation extraction dataset with a benchmark evaluation. In *ACM MM*, pages 4564–4573, 2023.
- [Kim *et al.*, 2020] Eun-Sol Kim, Woo-Young Kang, Kyoung-Woon On, Yu-Jung Heo, and Byoung-Tak Zhang. Hypergraph attention networks for multimodal learning. In *CVPR*, pages 14569–14578, 2020.
- [Kipf and Welling, 2016] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016.
- [Li *et al.*, 2019] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, 2019.
- [Li *et al.*, 2023] Qian Li, Shu Guo, Cheng Ji, Xutan Peng, Shiyao Cui, Jianxin Li, and Lihong Wang. Dual-gated fusion with prefix-tuning for multi-modal relation extraction. In *Findings of ACL*, pages 8982–8994, 2023.
- [Liang *et al.*, 2022] Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, and Fuchun Sun. Reasoning over different types of knowledge graphs: Static, temporal and multi-modal. *arXiv preprint arXiv:2212.05767*, 2022.
- [Liang *et al.*, 2023] Ke Liang, Yue Liu, Sihang Zhou, Wenxuan Tu, Yi Wen, Xihong Yang, Xiangjun Dong, and Xinwang Liu. Knowledge graph contrastive learning based on relation-symmetrical structure. *IEEE Transactions on Knowledge and Data Engineering*, 2023.
- [Liu *et al.*, 2024] Xiyang Liu, Chunming Hu, Richong Zhang, Kai Sun, Samuel Mensah, and Yongyi Mao. Multi-modal relation extraction via a mixture of hierarchical visual context learners. In *WWW*, pages 4283–4294, 2024.
- [Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

- [Lu *et al.*, 2022] Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. Unified structure generation for universal information extraction. In *ACL*, pages 5755–5772, 2022.
- [Petrov, 2022] Valentin V Petrov. Sums of independent random variables. In *Sums of Independent Random Variables*. 2022.
- [Redmon and Farhadi, 2018] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.
- [Scarselli *et al.*, 2008] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.
- [Soares *et al.*, 2019] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *ACL*, pages 2895–2905, 2019.
- [Tang *et al.*, 2020] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, pages 3713–3722, 2020.
- [Velickovic *et al.*, 2018] Petar Velickovic, Guillem Cucurull, Arantxa Casanova ands Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.
- [Wang *et al.*, 2022] Min Wang, Hao Yang, and Qing Cheng. GCL: graph calibration loss for trustworthy graph neural network. In *ACM MM*, pages 988–996, 2022.
- [Wu *et al.*, 2020a] Xiangping Wu, Qingcai Chen, Wei Li, Yulun Xiao, and Baotian Hu. Adahgnn: Adaptive hypergraph neural networks for multi-label image classification. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *ACM MM*, pages 284–293, 2020.
- [Wu *et al.*, 2020b] 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 *ACM MM*, pages 1038–1046, 2020.
- [Wu *et al.*, 2020c] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *TNNLS*, 32(1):4–24, 2020.
- [Xue *et al.*, 2022] Fuzhao Xue, Aixin Sun, Hao Zhang, Jinjie Ni, and Eng-Siong Chng. An embarrassingly simple model for dialogue relation extraction. In *ICASSP*, pages 6707–6711, 2022.
- [Zeng *et al.*, 2014] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344, 2014.
- [Zeng *et al.*, 2015] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *EMNLP*, pages 1753–1762, 2015.
- [Zeng *et al.*, 2023] Yawen Zeng, Qin Jin, Tengfei Bao, and Wenfeng Li. Multi-modal knowledge hypergraph for diverse image retrieval. In *AAAI*, pages 3376–3383, 2023.
- [Zhang *et al.*, 2017] Tongtao Zhang, Spencer Whitehead, Hanwang Zhang, Hongzhi Li, Joseph G. Ellis, Lifu Huang, Wei Liu, Heng Ji, and Shih-Fu Chang. Improving event extraction via multimodal integration. In *ACM MM*, pages 270–278, 2017.
- [Zhang *et al.*, 2018] Zizhao Zhang, Haojie Lin, Junjie Zhu, Xibin Zhao, and Yue Gao. Cross diffusion on multi-hypergraph for multi-modal 3d object recognition. In *PCM*, volume 11164, pages 38–49, 2018.
- [Zhang *et al.*, 2019] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. Heterogeneous graph neural network. In *KDD*, pages 793–803, 2019.
- [Zhang *et al.*, 2024] Zefan Zhang, Weiqi Zhang, Yanhui Li, and Tian Bai. Caption-aware multimodal relation extraction with mutual information maximization. In *ACM MM*, pages 1148–1157, 2024.
- [Zheng *et al.*, 2021a] Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. Multimodal relation extraction with efficient graph alignment. In *ACM MM*, pages 5298–5306, 2021.
- [Zheng *et al.*, 2021b] 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 *ICME*, pages 1–6, 2021.
- [Zheng *et al.*, 2023] Changmeng Zheng, Junhao Feng, Yi Cai, Xiaoyong Wei, and Qing Li. Rethinking multimodal entity and relation extraction from a translation point of view. In *ACL*, pages 6810–6824, 2023.
- [Zhong *et al.*, 2023] Fangming Zhong, Chenglong Chu, Zijie Zhu, and Zhikui Chen. Hypergraph-enhanced hashing for unsupervised cross-modal retrieval via robust similarity guidance. In *ACM MM*, pages 3517–3527, 2023.
- [Zhou *et al.*, 2020] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.
- [Zhu *et al.*, 2019] Dingyuan Zhu, Ziwei Zhang, Peng Cui, and Wenwu Zhu. Robust graph convolutional networks against adversarial attacks. In *KDD*, pages 1399–1407, 2019.