Prototype-Guided Multimodal Relation Extraction based on Entity Attributes

Zefan Zhang [1], Weiqi Zhang [1], Yanhui Li [1], Tian Bai [1*]

[1]College of Computer Science and Technology, Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, Jilin University

zefan23@mails.jlu.edu.cn, zwq23@mails.jlu.edu.cn, yanhui23@mails.jlu.edu.cn, baitian@jlu.edu.cn

Abstract

Multimodal Relation Extraction (MRE) aims to predict relations between head and tail entities based on the context of sentence-image pairs. Most existing MRE methods progressively incorporate textual and visual inputs to dominate the learning process, assuming both contribute significantly to the task. However, the diverse visual appearances and text with ambiguous semantics contain less-informative contexts for the corresponding relation. To tackle these challenges, we highlight the importance of semantically invariant entity attributes that encompass fine-grained categories. Towards this, we propose a novel Prototype-Guided Multimodal Relation Extraction (PG-MRE) framework based on Entity Attributes. Specifically, we first generate detailed entity explanations using Large Language Models (LLMs) to supplement the attribute semantics. Then, the Attribute Prototype Module (APM) refines attribute categories and condenses scattered entity attribute features into cluster-level prototypes. Furthermore, prototype-aligned attribute features guide diverse visual appearance features to produce compact and distinctive multimodal representations in the Relation Prototype Module (RPM). Extensive experiments demonstrate that our method gains superior relation classification capability (especially in scenarios involving various unseen entities), achieving new state-of-the-art performances on MNRE dataset.

Introduction

Multimodal Relation Extraction (MRE) (Zheng et al. 2021a,b) aims to identify the relation between the given entities based on image and text. It is a fundamental task in the interaction field of computer vision (Wu et al. 2024a,b; Liu et al. 2022; Yao, Li, and Xiao 2024; Yao et al. 2024) and natural language processing (Zhang, Ji, and Liu 2023; Zhang et al. 2024). Multimodal relation extraction aims to enhance computer understanding of textual relations by integrating visual data.

Previous research has primarily focused on identifying relations between entities from textual contexts (Hu et al. 2020b,a, 2021, 2022b). Other studies (Zheng et al. 2021a,b) have emphasized multimodal entity alignment, which involves locating textual entities within visual features to better understand their relations. Despite their remarkable efforts (Chen et al. 2022a; Wang et al. 2022b; Hu et al. 2023b), current research still faces several challenges:

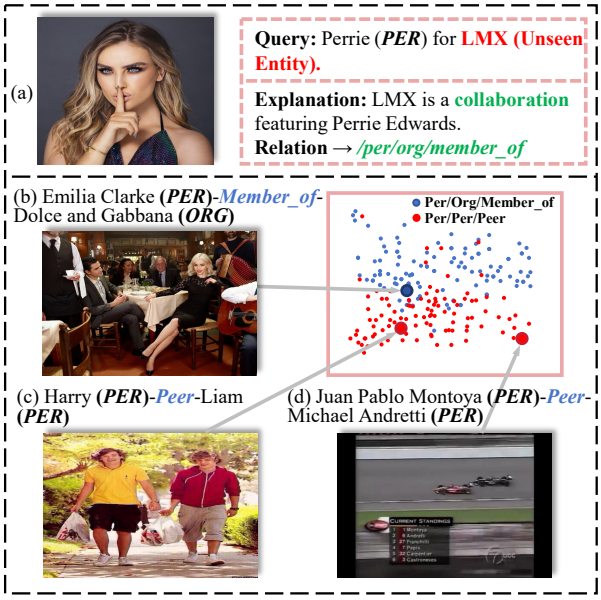

Figure 1: (a) shows the incomprehensible text (17.04%). (b), (c) and (d) describe the high intra-class variation (12.08%) and inter-class similarity (43.06%) in the diverse visual appearance. All ratios are the results we annotated on the MNRE test dataset.

1. Less-informative input contexts. In textual contexts, the semantics of input text is often too obscure. Entities may incorporate novel vocabularies unseen in both the training corpus and the pre-trained language model's vocabulary, as the "LMX" shown in Figure 1 (a). In visual contexts, entities exhibit diverse visual appearances. As the example shown in Figure 1 (c), images depicting the "peer" relation often feature two or more people. However, there is an inter-class similarity in relations, as seen in Figure 1 (b), where "member_of" and "peer" relations have similar visual representations. Additionally, there is significant visual variation within the same class, as demonstrated in Figure 1 (c) and (d), which shows a racing scene that differs greatly from typical "peer" scenarios. **Both diverse visual appearances and**

*The corresponding author.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

text with ambiguous semantics contain less-informative contexts for the corresponding relation and restrict extracting compact and distinctive multimodal representations. TMR (Zheng et al. 2023a) transfers knowledge in generative diffusion models as back-translation for MRE. CMG (Wu et al. 2023a) proposes to perform topic modeling over the input image and text, incorporating latent multimodal topic features to enrich the contexts. However, both text and image inputs often contain obscure semantics, which limits their performance in knowledge inheriting. Our key intuition is that **semantically invariant entity attributes with fine-grained categories are more reliable than visual and textual contexts in multimodal representations**.

2. Coarse-granularity attribute categories. Well-defined and comprehensive categories of entity attributes can facilitate more accurate entity relation classification. For example, there is only a single "nationality" relation between "person" and "country". However, the current dataset has limited categories of entity attributes (including person, organization, localization, and miscellaneous), often using "miscellaneous" to cover a large portion of entity attributes (such as country, religion, movies...). **These coarse-granularity attribute categories cannot alleviate the large intra-class variations within some categories which limits the effect of attribute information.**

In light of these challenges, we propose a Prototype-Guided Multimodal Relation Extraction (PG-MRE) framework. Specifically, to address the insufficient understanding of entity attributes in the input contexts, we propose to introduce detailed entity explanations by Large Language Models (LLMs). LLMs could provide explanations for entities with ambiguous semantics based on their vast reserve of commonsense knowledge, as the explanation shown in Figure 1 (a). Then, we introduce an Attribute Prototype Module (APM) to refine the existing attribute categories and extract attribute category prototypes from detailed entity explanations for inheriting fine-granularity attributes. APM captures the essential characteristics of the attribute and stores them in a prototype bank. Moreover, during multimodal fusion, attribute features guide the diverse visual features and are fused into compact and distinguishable representations by the Relation Prototype Module (RPM). Specifically, RPM captures the representative embeddings for each relation as relation prototypes and stores them in the corresponding prototype bank. During the inference phase, attribute and relation prototypes will be utilized to aid in classification.

In summary, our contributions are shown as follows:

- We propose a novel Prototype-Guided Multimodal Relation Extraction (PG-MRE) framework based on Entity Attributes to handle more general and flexible scenes.
- We introduce the Attribute Prototype Module which could expand the attribute categories and condense the scattered attribute semantics of entities to cluster-level prototypes based on the detailed entity explanations generated by LLMs. The Relation Prototype Module could produce compact and distinctive relation representations based on attribute features.
- We evaluate our methods on the MNRE dataset and demonstrate their superiority and generalization compared to previous state-of-the-art baselines.

Related Work

Multimodal Relation Extraction

As one of the key sub-tasks in the information extraction track (Wang et al. 2022a; Zheng et al. 2024; Lyu et al. 2023), Relation Extraction (RE) (Liu et al. 2019; Chen et al. 2022b) has recently garnered significant attention. Previous studies primarily focus on extracting relations from a single text modality (Hu et al. 2023c, 2022a). Recognizing that visual features from images can provide additional reasoning clues, researchers have proposed Multimodal Relation Extraction (MRE), which has subsequently gained more attention. Recent studies on MRE aim to leverage relevant images to enhance relation extraction.

In the early stages, numerous works (Lu et al. 2018; Moon, Neves, and Carvalho 2018; Zhang et al. 2018) propose encoding text using RNNs and images using CNNs, then establishing implicit interactions between the two modalities. Other works (Yu et al. 2020; Zhang et al. 2021) propose leveraging region-based image features to represent objects in images, exploiting fine-grained semantic correspondences based on Transformer. IFAformer (Li et al. 2023) uses a fine-grained multimodal alignment approach with Transformer, aligning visual and textual objects in the representation space. MoRe (Wang et al. 2022b) proposes retrieving textual evidence from a knowledge base constructed using Wikipedia. However, most methods overlook the issue of interference from irrelevant objects in images.

For multimodal alignment, RpBERT (Sun et al. 2021) learns a text-image similarity score to filter out irrelevant visual representations. HVPNeT (Chen et al. 2022a) proposes a visual prefix-guided fusion mechanism to remove irrelevant objects. For more fine-grained alignment, PROMU (Hu et al. 2023a) proposes entity-object and relation-image alignment pretraining tasks to improve MRE performance. HVFormer (Liu et al. 2024a) utilizes a novel two-stage hierarchical visual context fusion Transformer incorporating the mixture of multimodal experts framework to effectively represent and integrate hierarchical visual features into textual semantic representations. EEGA (Yuan et al. 2023) is an edge-enhanced graph alignment network to enhance the MRE task by aligning nodes and edges in the cross-graph.

Despite continuous progress, these methods overlook the paramount significance of the entity attributes and still face challenges due to the inaccurate semantic understanding of the contexts with ambiguous semantics.

Prototype Learning

Prototype Learning is a typical few-shot learning approach and has achieved success in recognizing classes with few training examples. Researchers (Chen et al. 2024; Zheng et al. 2023b) propose to model entities/predicates with prototype-aligned compact and distinctive representations and thereby establish matching between entity pairs and predicates in a common embedding space for relation recognition. CRUP (Liu et al. 2024b) stores the mean vectors of

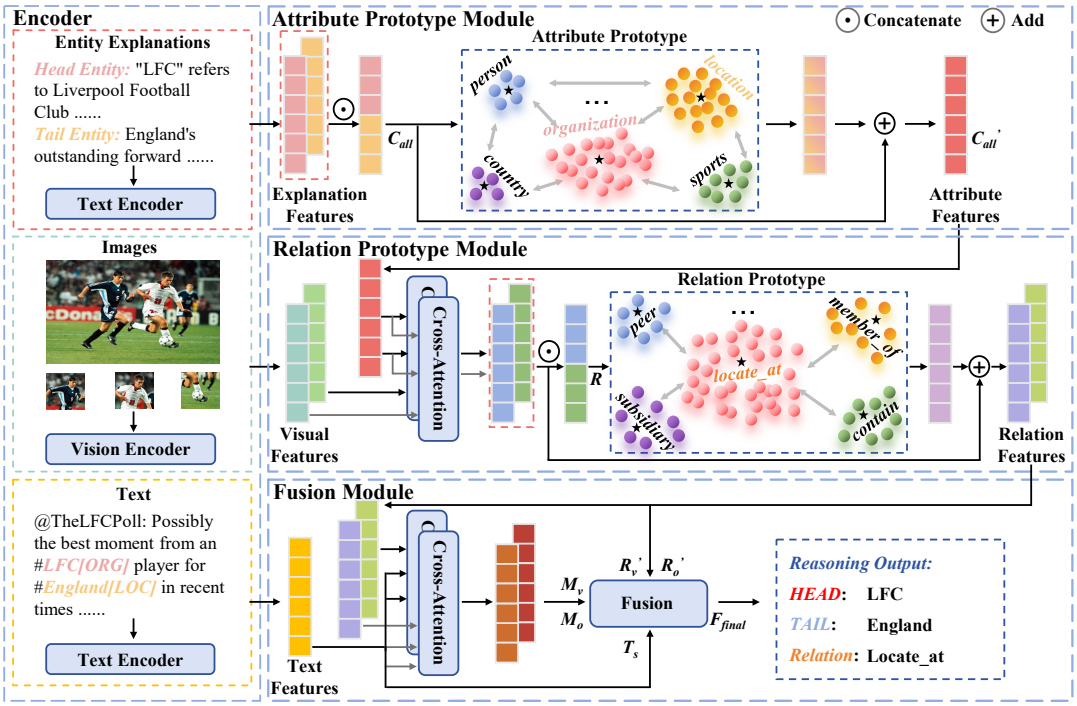

Figure 2: The framework of Prototype-Guided Multimodal Relation Extraction (PG-MRE).

representations belonging to new entity types as their prototypes and updates existing prototypes belonging to old types only based on representations of the current batch. The prototypes will be used for the nearest class mean classification. Researchers (Cheng et al. 2023) construct a sentence-wise prototype memory bank, enabling the network to focus on low-level visual and high-level clinical linguistic features.

In multimodal relation extraction, more accurate and sufficient entity attributes will provide great assistance for relation extraction. Prototype learning, with its characteristics of summarizing inter-class features (both textual and visual), is very suitable for learning entity attributes and relations in MRE. Therefore, we attempt for the first time to introduce prototype learning into MRE.

Method

Overview

Our Prototype-Guided Multimodal Relation Extraction (PG-MRE) framework is illustrated in Figure 2. Most previous methods fail to effectively utilize discriminative entity attributes and cannot handle contexts with ambiguous semantics. Our work addresses this problem in Multimodal Relation Extraction (MRE). Given an image-query pair, we first use Large Language Models (LLMs) to generate detailed entity explanations, supplementing the entity semantics. The Attribute Prototype Module then condenses these scattered entity semantics into cluster-level prototypes. Through the Relation Prototype Module, the attribute features and variable visual features are clustered into more compact and distinctive multimodal representations. Within the Attribute Prototype Module and Relation Prototype Module, we calculate the average features corresponding to each prototype and update the prototype features with momentum in each training iteration. During the test phase, we assign a label to the current features based on the nearest prototypes. These approaches ensure robust and discriminative classification, effectively harnessing the rich semantics and visual cues encapsulated within the prototypes.

Representation of Visual and Textual Features

Visual Features The image contains several visual objects linked to entities in the text, playing a significant role in aligning multimodal entities. Hence, we utilize object-level visual data provided by other works (Chen et al. 2022a). Additionally, global image features can convey abstract concepts for the whole relation between entities.

We begin by using the visual grounding toolkit to extract local visual objects with the top m salience (Zhang et al. 2021; Yang et al. 2019). Following other works (Chen et al. 2022a), we rescale both the original image and the object images to 224 × 224 pixels, referring to these as the original images $\mathcal{V}$ and the object images $\mathcal{O}$.

As illustrated in Figure 2, for the multimodal relation extraction task, we input the original images $\mathcal{V}$ and object images $\mathcal{O}$ into the Vision Transformer (Han et al. 2022) for encoding. This process obtains the original image features F_V and the object image features F_O. We define the image feature F_I to represent these uniformly, as shown in Eq.(1):

$$F_I = \{F_V, F_O\}. \tag{1}$$

Textual Features Since there are often contexts with ambiguous semantics and unseen entities in the training set, the model cannot accurately capture their semantics for relation prediction, such as "**Hola** (head entity) amigos. **Messerschmitt KR** (tail entity) 175 bubblecar (1953)." Therefore, we introduce large language models with a robust repository of commonsense knowledge to provide accurate explanations for the head and tail entities, C^H, C^T. We define the explanations C to represent these uniformly, as follows,

$$C = \left\{C^H, C^T\right\}. \tag{2}$$

The model can learn global semantics and achieve relation-level alignment with sentence-level features. Additionally, we leverage word-level features to ensure the model selects the correct features between different modalities. Hence, we leverage BERT (Devlin et al. 2018) to encode the input text T and entity explanations C. Then, we could obtain sentence-level features T_S, C_S and word-level features T_W, C_W where the sentence-level features are the pooler output of the word-level features.

Attribute Prototype Module

Current attribute categories are coarse-granularity and large intra-class variations within the same categories limit the effect of attribute information. Therefore, this module aims to capture the cluster-level attribute prototypes based on detailed entity explanations generated by LLMs.

We first define the attribute prototype bank as follows:

$$\mathcal{P}^a = \{p_i^a | i = 1...k\}, \tag{3}$$

where k is the number of attribute prototypes and each prototype is an embedding vector with d dimension. Then we concatenate the sentence-level head and tail entity explanation features and calculate the distance between textual features and prototypes,

$$C_{all} = MLP(C_S^H \odot C_S^T), \tag{4}$$

$$C_{all}{}' = p_j^a + C_{all}, \tag{5}$$

$$j = mindist(C_{all}, \mathcal{P}^a), \tag{6}$$

where $\odot$ denotes concatenation and MLP is the MLP (Multi-Layer Perceptron) layer. j is the assigned index and $dist(\cdot)$ is the euclidean distance. Here, the prototypes are discrete. The nearest neighbor searching method is not differentiable and can not be optimized with gradient backpropagation. Therefore, we follow (van den Oord, Vinyals, and kavukcuoglu 2017) and adopt their optimizing method by using the stopping gradient strategy. Specifically, it can be optimized as follows:

$$C_{all}{}' = sg[p_j - C_{all}] + C_{all}, \tag{7}$$

where $sg[\cdot]$ is the operation that stops the gradient on the parameters. The gradient for previous parameters is transmitted by the auxiliary term C_{all}, which makes the prototype bank trainable. In addition, we used momentum average optimization for prototype clustering. When the prototypes possess a non-zero count of features, the averages of those features are refreshed and serve as the updated prototypes. In general, this Attribute Prototype Module utilizes mini-batch deep learning clustering techniques to aggregate features that share similar attribute semantics.

Relation Prototype Module

The previous Attribute Prototype Module focuses on extracting textual entity attribute features. In this module, we focus on using the attribute features to guide the visual features and fusing them into compact and distinctive representations.

The relation-level multimodal fusion result encounters significant intra-class variation and severe inter-class similarity in previous studies. Subsequently, we establish multimodal features as relation-centric prototypes based on attribute features and use the key entity attribute features obtained from the Attribute Prototype Module to effectively guide the model to precisely comprehend and represent the intricate relations between entities. Specifically, we first use the cross-attention method to align and fuse entity attribute features with visual features, as follows:

$$CrossAtt\left(\mathcal{Q}, \mathcal{K}, \mathcal{V}\right) = softmax\left(\frac{\mathcal{Q}\mathcal{K}}{\sqrt{d}}\right)\mathcal{V}, \tag{8}$$

$$R_V = CrossAtt\left(F_V, C_{all}{}', C_{all}{}'\right), \tag{9}$$

$$R_O = CrossAtt\left(F_O, C_{all}{}', C_{all}{}'\right), \tag{10}$$

$$R = R_V \odot R_O. \tag{11}$$

The R denotes the dual-level multimodal fused result. It then transposes to the relation prototypes to capture compact relation-level features.

$$\mathcal{P}^r = \{p_i^r | i = 1...t\}, \tag{12}$$

$$R' = p_q^r + R, \tag{13}$$

$$q = mindist(R, \mathcal{P}^r), \tag{14}$$

where t is the number of relation prototypes. Finally, we perform residual links with the input visual features.

$$R_V{}' = R_V + R', \tag{15}$$

$$R_O{}' = R_O + R'. \tag{16}$$

Fusion Module

We first perform cross-attention calculations on the relation prototype enhanced visual features $R_V{}'$ and $R_O{}'$ with the word-level text feature T_W as follows:

$$M_V = CrossAtt\left(R_V{}', T_W, T_W\right), \tag{17}$$

$$M_O = CrossAtt\left(R_O{}', T_W, T_W\right). \tag{18}$$

Then, We add the residual link with M_V and M_O. Then fuse these two-level features and obtain M,

$$M = M_V + R_V{}' + M_O + R_O{}'. \tag{19}$$

Finally, we concatenate M and sentence-level text feature T_S through the MLP layers MLP_{final} and obtain the final fusion feature F_{final}:

$$F_{final} = MLP_{final}\left(M \odot T_S\right). \tag{20}$$

Methods	Accuracy	Precision	Recall	F1
Text Models				
BERT (2019)	74.42	58.58	60.25	59.40
PCNN (2015)	73.15	62.85	49.69	55.49
MTB (2019)	75.69	64.46	57.81	60.86
Text+ Image Models				
MoRe (2022)	79.87	65.25	67.32	66.27
MEGA (2021)	80.05	64.51	68.44	66.41
IFAformer (2023)	92.38	82.59	80.78	81.67
HVPNeT (2022)	92.52	83.64	80.78	81.85
TSVFN (2023)	92.67	85.16	82.07	83.02
MMIB (2024)	-	83.49	82.97	83.23
HVFormer (2024)	-	84.14	82.65	83.39
MRE-ISE (2023)	94.06	84.69	83.38	84.03
MRE (2023)	93.54	85.03	84.25	84.64
PROMU (2023)	-	84.95	85.76	84.86
TMR (2023)	-	90.48	87.66	89.05
PG-MRE (ours)	**96.34**	**92.41**	**91.25**	**91.82**

Table 1: Accuracy (%) comparison on MNRE testing set.

Classifier

The primary objective of the multimodal relation extraction task is to accurately predict the relation r that exists between the head and tail entities, based on the given set of labels L. To achieve this, we employ a dedicated [CLS] (Classification) head, which serves as a pivotal point for aggregating the probability distribution over the set of relation labels L with the softmax function. Finally, we calculate the RE loss with the cross-entropy loss function:

$$p(r|X) = softmax(X), \tag{21}$$

$$\mathcal{L}re = -\sum_{i=1}^{n} \log\left(p\left(r|F_{final}\right)\right). \tag{22}$$

Experiments

In this section, we conduct comprehensive experiments on the MNRE dataset. Our overarching goal is to answer the following research questions:

- RQ1: How does our model compare with the state-of-the-art MRE approaches on quantitative results?
- RQ2: How much do various components of our model contribute to its overall performance?
- RQ3: How does the proposed model compare against existing MRE methods in some special scenarios?
- RQ4: How does the proposed model compare to the state-of-the-art MRE methods on visual results?

Methods	Accuracy	Precision	Recall	F1
Base	94.61	90.08	86.56	88.29
Base+Ent	95.85	91.11	89.69	90.39
Base+Ent+APM	95.97	91.57	90.00	90.78
Base+Ent+RPM	95.97	91.22	90.94	91.08
Base+Ent+RPM+APM	**96.34**	**92.41**	**91.25**	**91.82**

Table 2: Ablation study on MNRE testing set. Base means baseline, Ent means incorporating the detailed entity explanations. RPM means the Relation Prototype Module, and APM denotes the Attribute Prototype Module.

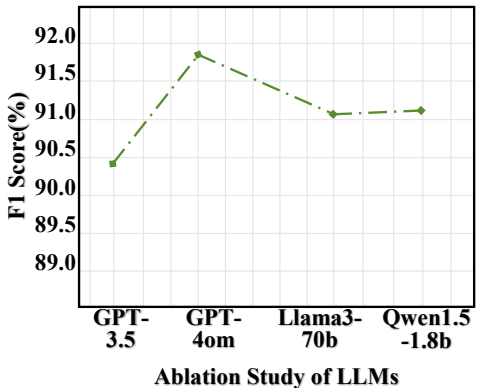

Figure 3: Ablation study of the explanations generated by different LLMs.

Datasets

We evaluate the model on MNRE dataset (Zheng et al. 2021a), which contains 15485 entity pairs, 9,201 text-image pairs, and 23 relation types. Successful extraction of the relation between two entities is achieved when the predicted relation type matches the ground truth. To assess the performance, we employ Accuracy, Precision, Recall, and F1 score as our evaluation metrics.

Implementation Details

For the text encoder of our model, we leverage the BERT-Base default tokenizer with a max length of 128 to preprocess data. For the vision encoder, we leverage the vision Transformer (Han et al. 2022) to encode the original images and object images. The batch size is 16 and the learning rate is 2e-5. We train the model for 8 epochs. Moreover, the dimension of the hidden states d is set to 768. we set the entity explanation length to 40 and uniformly use a bert-base-uncased encoding. All optimizations are performed with the AdamW optimizer with a linear warmup of learning rate over the first 10% of gradient updates to a maximum value, then linear decay over the remainder of the training. And weight decay on all non-biased parameters is set to 0.01. We set the number of image objects l to 3.

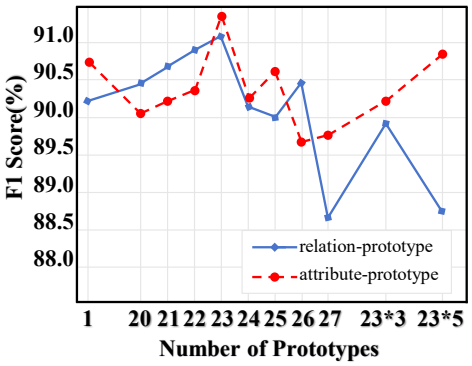

Figure 4: Ablation study of the different number of prototypes in APM and RPM.

Methods	Accuracy (%)	
	MNRE_US	MNRE_UT
HVPNet (2022)	87.43	86.94
TMR (2023)	93.71	94.49
PG-MRE (ours)	**95.81**	**96.00**

Table 3: Comparison results between HVPNeT (Chen et al. 2022a), TMR (Zheng et al. 2023a), and our PG-MRE on MNRE_US and MNRE_UT.

Baselines

We compare our model with the following baselines for a comprehensive comparison.

BERT (Devlin et al. 2018), **PCNN** (Zeng et al. 2015) and **MTB** (Soares et al. 2019) are Text-based RE methods. **MoRe** (Wang et al. 2022b) injects knowledge-aware information into multimodal studies using multimodal retrieval. **MEGA** (Zheng et al. 2021a), **IFAformer** (Li et al. 2023), and **MMIB** (Cui et al. 2024) are multimodal alignment based models. **HVPNeT** (Chen et al. 2022a) treats visual representations as visual prefixes that can be inserted to guide textual representations of error-insensitive prediction decisions. **TSVFN** (Zhao, Gao, and Guo 2023) combines the powerful modeling capabilities of graph neural networks and Transformer networks to fully fuse critical information between visual and textual modalities. **MRE** (Hu et al. 2023b) uses cross-modal retrieval for obtaining multimodal evidence to improve prediction accuracy and synthesize visual and textual information for relational reasoning. **MRE-ISE** (Wu et al. 2023b) introduces a novel idea of simultaneous information subtraction and addition for multimodal relation extraction. **PROMU** (Hu et al. 2023a) enables the extraction of self-supervised signals from massive unlabeled image-caption pairs to pretrain multimodal fusion modules. **TMR** (Zheng et al. 2023a) implements multimodal versions of back-translation and high-resource bridging, which provide a multi-view to the misalignment between modalities. **HVFormer** (Liu et al. 2024a) proposes a novel two-stage hierarchical visual context fusion Transformer incorporating the mixture of multimodal experts framework.

Quantitative Comparison with State-of-the-art Methods (RQ1)

The main results are shown in Table. 1. It could be seen that our model achieves the SOTA results.

Obviously, our method significantly outperforms text-based approaches. In comparison with multimodal alignment-based methods (**MEGA** (Zheng et al. 2021a), **IFAformer** (Li et al. 2023), and **MMIB** (Cui et al. 2024)), our method also achieves superior performance, which indicates that previous multimodal alignment strategies overlook the entities with ambiguous semantics. The current state-of-the-art (SOTA) model is **TMR** (Zheng et al. 2023a), which leverages the diffusion strategy to generate a large amount of additional image data to enhance the understanding of relations between entities. However, due to the crucial guiding role of text in multimodal alignment, the neglect of entity attributes still renders its performance inferior to ours.

Therefore, the results indicate that our method provides more accurate entity semantics and clusters more efficient attribute and relation prototypes. PG-MRE achieves stable and excellent results (Accuracy +2.28%, Precision +1.93%, Recall +3.59%, F1 +2.77%) which also indicates that our PG-MRE has advantages in identifying diverse visual appearances and contexts with ambiguous semantics.

Ablation Study (RQ2)

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

Explanations generated by different LLMs In Table 2, comparing the results "Base" and "Base+Ent", we could find that the detailed entity explanations lead to a steady improvement (Accuracy +1.24%, Precision +1.03%, Recall +3.13%, F1 +2.1%). Among them, the metric of Recall is improved significantly, which also means that entity explanations play an important role in obscure context understanding. Additionally, we evaluate the generated explanation's quality and find that only 3.34% of the explanations contain noisy tokens, with minimal impact on the model.

In Figure 3, we conduct an ablation study over the entity explanations generated by different Large Language Models (LLMs), such as GPT-3.5, GPT-4o mini, Llama3-70B (Touvron et al. 2023), and Qwen1.5-1.8B (Bai et al. 2023). Good explanations require a vast reserve of commonsense knowledge. The results show that current LLMs can provide sufficient explanations, such as "MLB, or Major League Baseball, is a professional baseball organization...". Their performances fluctuate within the normal range, and the model with the explanations generated by GPT-4o mini achieves the current best performance.

Attribute Prototype Module In Table 2, comparing the results "Base+Ent" and "Base+Ent+APM", we could find that the APM can boost model performance. It indicates that the APM could assist the model enriches the attribute cate-

Case 1	Case 2	Case 3
Text: @_PLICE: A lot of people don't know "My Prince" from Mr. Bones is John Diggle in Arrow.	**Text**: @paddypope:When Arsenal go 4th but our next game is City...	**Text**: @CLS_Rob: Was looking for a title for my @CAPhys talk when I spotted this on the @CanLightSource control room.
Head : Mr.Bones *Attribute:* ***PER*** *Tail :* Arrow *Attribute:* ***ORG***	*Head :* Arsenal *Attribute:* ***ORG*** *Tail :* City *Attribute:* ***ORG***	*Head :* @CAPhys *Attribute:* ***ORG*** *Tail :* @CanLightSource *Attribute:* ***ORG***
Head Entity Explanation: Mr. Bones is a fictional character who is depicted as ..... *Tail Entity Explanation:* Arrow is an American crime drama television series set in the	*Head Entity Explanation:* The entity "Arsenal" refers to the professional football club ..... *Tail Entity Explanation:* ..... Manchester City (also an English Premier League club)	*Head Entity Explanation:* The entity "@CAPhys" is likely a name given to a person or organization ..... *Tail Entity Explanation:* @CanLightSource" entity represents a specific platform .....
Gold Relation: ***per/org/member_of***	*Gold Relation:* ***org/org/subsidiary***	*Gold Relation:* ***/org/org/alternate_names***
HVPNeT: per/per/peer ✗	per/loc/place_of_residence ✗	None ✗
TMR: per/per/alternate_names ✗	per/loc/place_of_residence ✗	None ✗
PG-MRE: per/org/member_of ✓	org/org/subsidiary ✓	org/org/alternate_names ✓

Figure 5: Several cases comparison predicted by HVPNeT (Chen et al. 2022a), TMR (Zheng et al. 2023a), and our PR-MRE.

gories, and the cluster-level attribute prototypes facilitate the accurate reasoning process.

In addition, in Figure 4, we conduct an ablation study over the attribute prototype number, and it could be found that 23 is the best number.

Relation Prototype Module In Table 2, comparing the results "Base+Ent", "Base+Ent+APM" and "Base+Ent+RPM", we could find that the model with RPM has a steady improvement. However, the lack of the APM results in not concise and accurate attribute semantics, thereby interfering with the RPM. Similarly, not incorporating the RPM after the APM leads to interference from related relation features during multimodal fusion. The model achieves the best performance only when both work together. In addition, in Figure 4, we also conduct an ablation study over the relation prototype number. It could be found that 23 is the best number which is the same as the number of relation type.

Comparison Results in Some Special Scenarios (RQ3)

To test the robustness and the generalization of our model, we experiment with it in some special scenarios.

In the test set of the MNRE dataset, we select samples containing entities not seen in the training set as unseen samples, splitting them as a subset called MNRE_US. The results show that 76.89% (1241/1614) of the samples in the test set contained entities not seen in the training set. Additionally, we used BERT as the text encoder, but the dataset included words outside BERT's vocabulary that could not be tokenized. We split test set samples containing these untokenizable entities as a subset called MNRE_UT. The results showed that 69.76% (1126/1614) of the samples in the test set contained untokenizable entities.

We test HVPNeT, TMR, and our PG-MRE on these two data subsets. As shown in the Table 3, the results indicate that PG-MRE consistently improves accuracy (MNRE_US +2.10%, MNRE_UT +1.51%). These experimental results demonstrate that our method maintains stable performance on unseen and untokenizable entities.

Case Analysis (RQ4)

We conduct a case study of PG-MRE with the current typical works (HVPNeT (Chen et al. 2022a) and TMR (Zheng et al. 2023a)), as shown in Figure 5.

In case 1, HVPNeT (Chen et al. 2022a) and TMR (Zheng et al. 2023a) can not understand the "Arrow", they may focus on the image for supplement information. The image describes two men, and HVPNeT tends to choose "peer". However, our PG-MRE provides an accurate explanation of "Arrow" and chooses the right relation.

Similarly, in case 2, the entity of "City" has uncommon semantics. Its real meaning is "Manchester City which is an English Premier League club". However, TMR and HVPNeT tend to think it is "LOC" and choose the wrong relation without explanations. In case 3, these entities are unseen and untokenizable. Their ambiguous contexts make it difficult to determine the relations between entities directly. After obtaining explanations for key entities, further summarizing and clustering their attribute information based on the critical words ("organization", "platform") enables our model to understand and classify the right relations accurately.

Conclusion

In this work, we propose a Prototype-Guided Multimodal Relation Extraction (PG-MRE) network. We emphasize the significant importance of entity attributes, and we incorporate detailed entity explanations for understanding diverse visual appearances and text with ambiguous semantics. Then, we utilize the Attribute Prototype Module to expand the attribute categories and condense the scattered attribute semantics of entities to cluster-level prototypes. The Relation Prototype Module is used to produce compact and distinctive entity relation representations based on the attribute features. Extensive experiments show that our model could achieve the state-of-the-art (SOTA) result. In the future, we intend to explore the Retrieval Augmented Generation (RAG) multimodal relation extraction models to provide more real-time entity explanations. We are also interested in adapting our approach to more general scenarios and open-vocabulary scenarios.

Acknowledgments

This work is supported by the National Natural Science Foundation of China [U21A20390], the Development Project of Jilin Province of China [20240601039RC] and the Fundamental Research Funds for the Central University, JLU.

References

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Chen, L.; Song, Y.; Cai, Y.; Lu, J.; Li, Y.; Xie, Y.; Wang, C.; and He, G. 2024. Multi-Prototype Space Learning for Commonsense-Based Scene Graph Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(2): 1129–1137.

Chen, X.; Zhang, N.; Li, L.; Yao, Y.; Deng, S.; Tan, C.; Huang, F.; Si, L.; and Chen, H. 2022a. Good visual guidance makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. *arXiv preprint arXiv:2205.03521*.

Chen, X.; Zhang, N.; Xie, X.; Deng, S.; Yao, Y.; Tan, C.; Huang, F.; Si, L.; and Chen, H. 2022b. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web conference 2022*, 2778–2788.

Cheng, P.; Lin, L.; Lyu, J.; Huang, Y.; Luo, W.; and Tang, X. 2023. Prior: Prototype representation joint learning from medical images and reports. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21361–21371.

Cui, S.; Cao, J.; Cong, X.; Sheng, J.; Li, Q.; Liu, T.; and Shi, J. 2024. Enhancing multimodal entity and relation extraction with variational information bottleneck. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1): 87–110.

Hu, X.; Chen, J.; Liu, A.; Meng, S.; Wen, L.; and Yu, P. S. 2023a. Prompt me up: Unleashing the power of alignments for multimodal entity and relation extraction. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5185–5194.

Hu, X.; Guo, Z.; Teng, Z.; King, I.; and Yu, P. S. 2023b. Multimodal relation extraction with cross-modal retrieval and synthesis. *arXiv preprint arXiv:2305.16166*.

Hu, X.; Hong, Z.; Zhang, C.; King, I.; and Yu, P. 2023c. Think rationally about what you see: Continuous rationale extraction for relation extraction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2436–2440.

Hu, X.; Jiang, Y.; Liu, A.; Huang, Z.; Xie, P.; Huang, F.; Wen, L.; and Yu, P. S. 2022a. Entity-to-text based data augmentation for various named entity recognition tasks. *arXiv preprint arXiv:2210.10343*.

Hu, X.; Liu, S.; Zhang, C.; Li, S.; Wen, L.; and Yu, P. S. 2022b. Hiure: Hierarchical exemplar contrastive learning for unsupervised relation extraction. *arXiv preprint arXiv:2205.02225*.

Hu, X.; Zhang, C.; Ma, F.; Liu, C.; Wen, L.; and Yu, P. S. 2020a. Semi-supervised relation extraction via incremental meta self-training. *arXiv preprint arXiv:2010.16410*.

Hu, X.; Zhang, C.; Xu, Y.; Wen, L.; and Yu, P. S. 2020b. SelfORE: Self-supervised relational feature learning for open relation extraction. *arXiv preprint arXiv:2004.02438*.

Hu, X.; Zhang, C.; Yang, Y.; Li, X.; Lin, L.; Wen, L.; and Yu, P. S. 2021. Gradient imitation reinforcement learning for low resource relation extraction. *arXiv preprint arXiv:2109.06415*.

Li, L.; Chen, X.; Qiao, S.; Xiong, F.; Chen, H.; and Zhang, N. 2023. On analyzing the role of image for visual-enhanced relation extraction (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 16254–16255.

Liu, X.; Hu, C.; Zhang, R.; Sun, K.; Mensah, S.; and Mao, Y. 2024a. Multimodal Relation Extraction via a Mixture of Hierarchical Visual Context Learners. In *Proceedings of the ACM on Web Conference 2024*, 4283–4294.

Liu, Y.; Meng, F.; Zhang, J.; Xu, J.; Chen, Y.; and Zhou, J. 2019. GCDT: A global context enhanced deep transition architecture for sequence labeling. *arXiv preprint arXiv:1906.02437*.

Liu, Y.; Wang, P.; Ke, W.; Li, G.; Chen, X.; Zhao, J.; and Shang, Z. 2024b. Unify Named Entity Recognition Scenarios via Contrastive Real-Time Updating Prototype. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14035–14043.

Liu, Z.; Wu, S.; Xu, C.; Wang, X.; Zhu, L.; Wu, S.; and Feng, F. 2022. Copy motion from one to another: Fake motion video generation. *arXiv preprint arXiv:2205.01373*.

Lu, D.; Neves, L.; Carvalho, V.; Zhang, N.; and Ji, H. 2018. 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)*, 1990–1999.

Lyu, X.; Gao, L.; Zeng, P.; Shen, H. T.; and Song, J. 2023. Adaptive Fine-Grained Predicates Learning for Scene Graph Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 13921–13940.

Moon, S.; Neves, L.; and Carvalho, V. 2018. Multimodal named entity recognition for short social media posts. *arXiv preprint arXiv:1802.07862*.

Soares, L. B.; FitzGerald, N.; Ling, J.; and Kwiatkowski, T. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.

Sun, L.; Wang, J.; Zhang, K.; Su, Y.; and Weng, F. 2021. RpBERT: a text-image relation propagation-based BERT

model for multimodal NER. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 13860–13868.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

van den Oord, A.; Vinyals, O.; and kavukcuoglu, k. 2017. Neural Discrete Representation Learning. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, S.; Gao, L.; Lyu, X.; Guo, Y.; Zeng, P.; and Song, J. 2022a. Dynamic Scene Graph Generation via Temporal Prior Inference. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 5793–5801. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Wang, X.; Cai, J.; Jiang, Y.; Xie, P.; Tu, K.; and Lu, W. 2022b. Named entity and relation extraction with multimodal retrieval. *arXiv preprint arXiv:2212.01612*.

Wu, S.; Chen, H.; Yin, Y.; Hu, S.; Feng, R.; Jiao, Y.; Yang, Z.; and Liu, Z. 2024a. Joint-Motion Mutual Learning for Pose Estimation in Video. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8962–8971.

Wu, S.; Fei, H.; Cao, Y.; Bing, L.; and Chua, T.-S. 2023a. Information Screening whilst Exploiting! Multimodal Relation Extraction with Feature Denoising and Multimodal Topic Modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14734–14751.

Wu, S.; Fei, H.; Cao, Y.; Bing, L.; and Chua, T.-S. 2023b. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*.

Wu, S.; Liu, Z.; Zhang, B.; Zimmermann, R.; Ba, Z.; Zhang, X.; and Ren, K. 2024b. Do as I Do: Pose Guided Human Motion Copy. *IEEE Transactions on Dependable and Secure Computing*.

Yang, Z.; Gong, B.; Wang, L.; Huang, W.; Yu, D.; and Luo, J. 2019. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4683–4693.

Yao, J.; Lai, Y.; Kou, H.; Wu, T.; and Liu, R. 2024. QE-BEV: Query evolution for bird's eye view object detection in varied contexts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2927–2935.

Yao, J.; Li, C.; and Xiao, C. 2024. Swift Sampler: Efficient Learning of Sampler by 10 Parameters. *arXiv preprint arXiv:2410.05578*.

Yu, J.; Jiang, J.; Yang, L.; and Xia, R. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. Association for Computational Linguistics.

Yuan, L.; Cai, Y.; Wang, J.; and Li, Q. 2023. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 11051–11059.

Zeng, D.; Liu, K.; Chen, Y.; and Zhao, J. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 1753–1762.

Zhang, D.; Wei, S.; Li, S.; Wu, H.; Zhu, Q.; and Zhou, G. 2021. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14347–14355.

Zhang, Q.; Fu, J.; Liu, X.; and Huang, X. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhang, Z.; Ji, Y.; and Liu, C. 2023. Knowledge-Aware Causal Inference Network for Visual Dialog. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, ICMR '23, 253–261. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701788.

Zhang, Z.; Zhang, W.; Li, Y.; and Bai, T. 2024. Caption-Aware Multimodal Relation Extraction with Mutual Information Maximization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1148–1157.

Zhao, Q.; Gao, T.; and Guo, N. 2023. TSVFN: Two-stage visual fusion network for multimodal relation extraction. *Information Processing & Management*, 60(3): 103264.

Zheng, C.; Feng, J.; Cai, Y.; Wei, X.; and Li, Q. 2023a. 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 (Volume 1: Long Papers)*, 6810–6824.

Zheng, C.; Feng, J.; Fu, Z.; Cai, Y.; Li, Q.; and Wang, T. 2021a. Multimodal relation extraction with efficient graph alignment. In *Proceedings of the 29th ACM international conference on multimedia*, 5298–5306.

Zheng, C.; Gao, L.; Lyu, X.; Zeng, P.; El Saddik, A.; and Shen, H. T. 2024. Dual-Branch Hybrid Learning Network for Unbiased Scene Graph Generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3): 1743–1756.

Zheng, C.; Lyu, X.; Gao, L.; Dai, B.; and Song, J. 2023b. Prototype-based embedding network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22783–22792.

Zheng, C.; Wu, Z.; Feng, J.; Fu, Z.; and Cai, Y. 2021b. 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)*, 1–6. IEEE.