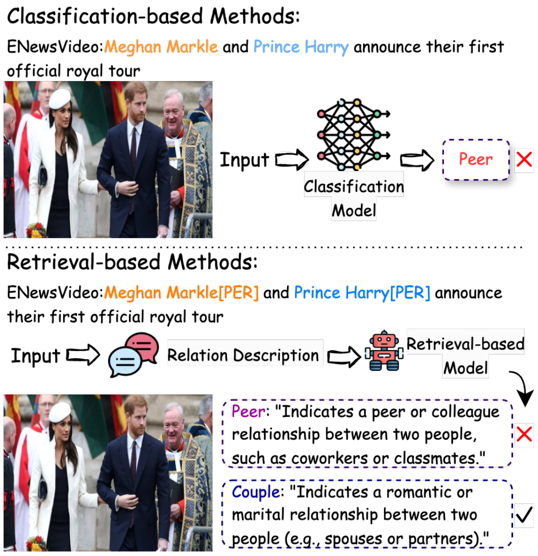
Retrieval over Classification: Integrating Relation Semantics for Multimodal Relation Extraction

Lei Hei[*], Tingjing Liao[*], Yingxin Pei, Yiyang Qi, Jiaqi Wang, Ruiting Li, Feiliang Ren[†]
School of Computer Science and Engineering,
Northeastern University, Shenyang 110819, China
renfeiliang@cse.neu.edu.cn

Abstract

Relation extraction (RE) aims to identify semantic relations between entities in unstructured text. Although recent work extends traditional RE to multimodal scenarios, most approaches still adopt classification-based paradigms with fused multimodal features, representing relations as discrete labels. This paradigm has two significant limitations: (1) it overlooks structural constraints like entity types and positional cues, and (2) it lacks semantic expressiveness for fine-grained relation understanding. We propose Retrieval Over Classification (ROC), a novel framework that reformulates multimodal RE as a retrieval task driven by relation semantics. ROC integrates entity type and positional information through a multimodal encoder, expands relation labels into natural language descriptions using a large language model, and aligns entity-relation pairs via semantic similarity-based contrastive learning. Experiments show that our method achieves state-of-the-art performance on the benchmark datasets MNRE and MORE and exhibits stronger robustness and interpretability.



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

ENewsVideo: Meghan Markle[PER] and Prince Harry[PER] announce their first official royal tour

Peer: "Indicates a peer or colleague relationship between two people, such as coworkers or classmates."

Couple: "Indicates a romantic or marital relationship between two people (e.g., spouses or partners)."

Figure 1: Comparison of classification-based and retrieval-based methods.

1 Introduction

Relation extraction (RE) is a fundamental task in information extraction, aiming to identify semantic relations between entities from unstructured texts automatically(Soares et al., 2019a; Yu et al., 2020a). It provides crucial structured data for downstream applications such as knowledge graph construction and question answering(Luo et al., 2018; Li et al., 2019b; Yu et al., 2020b).

However, traditional text-only RE methods face two key challenges. First, the inherent ambiguity of natural language often leads to incorrect predictions due to insufficient contextual information. Second, real-world data is frequently accompanied by visual information, and accurate relation inference in such scenarios usually requires joint reasoning over both textual and visual modalities(Zheng et al., 2021b). These limitations motivate research in multimodal relation extraction (MRE).

Existing MRE approaches primarily follow two paradigms: modality fusion(Chen et al., 2022c; Zhao et al., 2023; Liu et al., 2024b; Cui et al., 2024) and modality alignment(Zheng et al., 2021a; Wu et al., 2023; Hu et al., 2023; Li et al., 2024). Despite their technical differences, both paradigms ultimately rely on mapping multimodal features into a discrete set of predefined relation categories—a classification-based framework. This paradigm faces two significant limitations.

First, it neglects structural constraints such as entity types and positions. For instance, in a location_at relation, the subject is typically a location or an organization, while the object is usually a location. Without modeling such priors, the model

[*] Equal contribution.
[†] Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 18678–18693
November 4-9, 2025 ©2025 Association for Computational Linguistics

must search for relations across many irrelevant entity pairs, significantly increasing reasoning difficulty and reducing classification accuracy.

Second, fixed label indices limit the model's semantic expressiveness for fine-grained relation understanding. Figure 1 illustrates a concrete example: due to the semantic similarity between "Peer" and "Couple" in the representation space, a classification model tends to misclassify a married couple as "Peer". This mistake stems from the limited expressiveness of discrete labels, which fail to capture semantically similar but fundamentally different relations, weakening the model's ability to distinguish subtle relational nuances.

To address these limitations, we propose a novel framework: Retrieval Over Classification (ROC), which reformulates multimodal relation extraction as a semantically-driven retrieval task.

To address the first challenge, we incorporate entity types and positions to constrain the candidate relation space. We use the Stanford NER toolkit(Qi et al., 2020a) to identify entity types and embed them into the input as explicit semantic prompts, guiding the model to learn type-aware semantics. Entity positions help localize entity pairs and construct centered representations. We design a type-aware multimodal encoder to jointly encode these structural cues, effectively narrowing the candidate space and improving classification accuracy.

To address the second challenge, we replace discrete relation labels with natural language descriptions to enhance semantic expressiveness. Using GPT-4o(OpenAI et al., 2024), we generate descriptive sentences for each relation, followed by manual verification to ensure quality. A relational semantic encoder then transforms these descriptions into global semantic representations. Compared to fixed labels, this approach enables finer-grained relation modeling and improves the model's ability to distinguish semantically similar relations.

To align entity pairs with relation semantics, we introduce a contrastive retrieval strategy. The multimodal entity pair encoder and the relation semantic encoder project their features into a shared semantic space. The model is trained to maximize the similarity between matched pairs while suppressing irrelevant candidates, enabling accurate relation prediction. This retrieval-based paradigm integrates multimodal information, maintains semantic interpretability through natural language, and mitigates the label rigidity often observed in classification-based approaches.

Our contributions are summarized as follows:

- We propose ROC, reformulating multimodal relation extraction as a semantic retrieval task, offering a paradigm shift from traditional classification-based methods.
- We design a type-aware multimodal encoder incorporating entity type and position to effectively constrain the candidate relation space. Additionally, we construct natural language relation representations and introduce a relation semantic encoder to enhance fine-grained semantic modeling. A cross-modal contrastive retrieval mechanism aligns entity pairs with relation semantics in a shared space, enabling semantically consistent relation prediction.
- ROC achieves state-of-the-art performance on both the multimodal relation extraction benchmark MNRE(Zheng et al., 2021b) and the cross-modal dataset MORE(He et al., 2023), demonstrating the effectiveness and generalizability of our approach.

2 Related Work

Modality Fusion Paradigms Modality fusion methods aim to enable interactive learning of visual and textual features through deep neural networks. Representative approaches include HVPNeT (Chen et al., 2022b), which introduces multiscale visual features and leverages a dynamic gating mechanism to guide the language model in capturing image context. MMIB (Cui et al., 2024), built on variational autoencoders, incorporates mutual information maximization (Shannon, 1948) and the information bottleneck principle (Alemi et al., 2017) to alleviate representational discrepancies across modalities. While these methods effectively enhance cross-modal semantic perception, they fundamentally rely on mapping fused features to a predefined label space using discrete relation classifiers. This paradigm overlooks structural priors between entity types, making it challenging to model semantically continuous relationships, thus exhibiting a strong dependence on rigid labels.

Modality Alignment Paradigms Modality alignment methods introduce structured intermediate representations to guide semantic mapping across modalities, thereby enhancing relation modeling. For example, MEGA (Zheng et al., 2021a) aligns syntactic dependency structures from text with visual scene graphs (Tang et al., 2020) from images to improve entity-relation extraction accuracy. MREISE (Wu et al., 2023) leverages CLIP (Radford et al., 2021) to construct cross-modal graph structures and optimizes their representation via an information bottleneck mechanism. These methods improve model interpretability through explicit structural constraints, but their ability to model relations remains limited by the expressiveness and applicability of the prior structures. This limitation becomes more pronounced in diverse and complex contexts, where capturing the semantic dependencies between entity types and flexibly modeling structural constraints are challenging.

Emerging Paradigms In recent years, some studies have sought to overcome the limitations of traditional classification paradigms. EEGA (Yuan et al., 2023) introduces an end-to-end framework for joint extraction of entities and relations, removing the dependence on predefined entity labels. MOREformer (He et al., 2023) explores cross-modal object-entity relation modeling, offering greater adaptability in multimodal scenarios. However, these approaches still follow a two-stage pipeline of “feature learning followed by discrete classification,” and have yet to break through the representational bottleneck of classification-based semantic modeling fundamentally.

The above methods primarily focus on classification, emphasizing the extraction of textual and visual features while neglecting the semantic modeling of relationships themselves. Therefore, we propose a retrieval-based paradigm incorporating relational semantics which enables the model to extract multimodal features while constraining the search space for entity pairs through type constraints. It also leverages natural language descriptions to provide fine-grained semantic information, more effectively facilitating the model’s learning process.

3 Methodology

3.1 Task Definition

The task of MRE can be formally defined as follows: Given an input text sequence $T = [w_1, w_2, \ldots, w_n]$ and its associated image I, the goal is to predict a set of relational triples $Y = \{(s, r, o)_c\}_{c=1}^{C}$. Here, $s \in E$ and $o \in E$ denote the subject and object entities, where E is the set of all entities in the input. The relation $r \in R$ is selected from a predefined relation set R, and $(s, r, o)_c$ represents the c-th predicted relation triple.

Unlike traditional classification paradigms, we do not directly select a relation type r from the discrete relation label space R. Instead, each relation type is represented by a *natural language description*, which is encoded into a *shared semantic space* via a *relation semantic encoder*. Based on the fully integrated multimodal representation of the entity pair (s, o), the model retrieves the relation description that is most semantically aligned in this space to determine the relation type r. This reformulates multimodal relation extraction as a semantics-driven multimodal retrieval task.

3.2 Overview

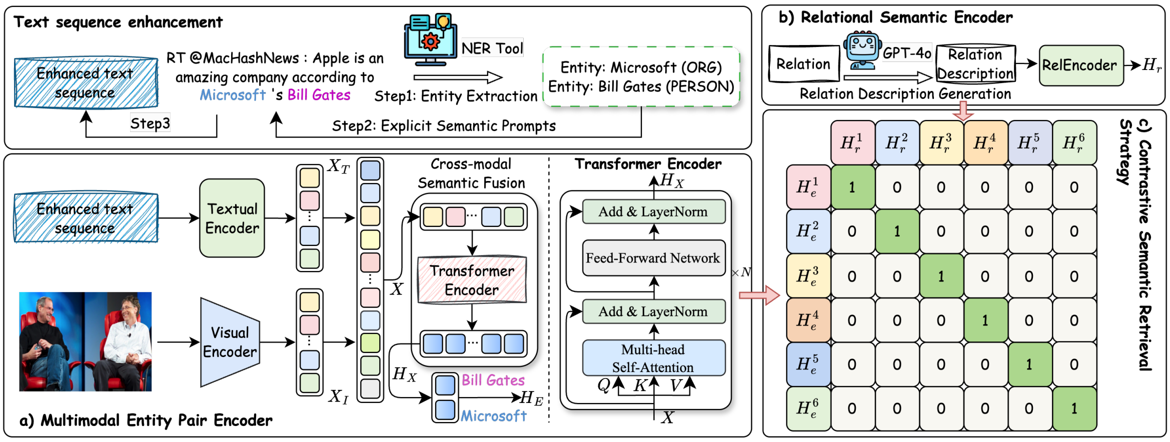
The overall architecture of our proposed ROC model is illustrated in Figure 2. It mainly consists of three core components: (1) **Multimodal Entity Pair Encoder**: It integrates explicit entity type annotations with a Transformer-based interaction mechanism to fuse textual and visual features, enhancing the semantic representation of entities in cross-modal contexts (Section 3.3); (2) **Relation Semantic Encoder**: Relation types are represented in natural language and encoded into a unified semantic space using an independent language model, which explicitly models the semantic differences between relation categories (Section 3.4); (3) **Contrastive Semantic Retrieval Strategy**: A matching mechanism is established between multimodal entity pair representations and relation semantics. The model achieves more discriminative relation extraction by optimizing the semantic similarity between each entity pair and its corresponding relation description (Section 3.5).

3.3 Multimodal Entity Pair Encoder

To constrain the candidate relation space and improve the accuracy of relation extraction, we design a type-aware multimodal entity pair encoder. It jointly models entity types and positional information to guide the model in filtering out semantically or spatially inconsistent entity combinations, while effectively integrating textual and visual features.

We use the Stanford NER tool(Qi et al., 2020b) to identify entity types in the input text. The recognized type information is embedded into the original text sequence as an explicit semantic prompt, guiding the model to perceive type priors during encoding. A pretrained BERT(Devlin et al., 2019) model encodes the enhanced textual sequence to obtain the textual feature representation X_T.

Meanwhile, we adopt a pretrained Vision Transformer (ViT)(Dosovitskiy et al., 2021) to extract


Text sequence enhancement
Enhanced text sequence
RT @MacHashNews : Apple is an amazing company according to Microsoft 's Bill Gates
NER Tool
Step1: Entity Extraction
Entity: Microsoft (ORG)
Entity: Bill Gates (PERSON)
Step3
Step2: Explicit Semantic Prompts
Enhanced text sequence
Textual Encoder
Visual Encoder
X_T
X_I
Cross-modal Semantic Fusion
X
Transformer Encoder
H_X
Bill Gates
Microsoft
H_E
Transformer Encoder
H_X
Add & LayerNorm
Feed-Forward Network
Add & LayerNorm
Multi-head Self-Attention
Q K V
X
$\times N$
a) Multimodal Entity Pair Encoder
b) Relational Semantic Encoder
Relation
GPT-4o
Relation Description
RelEncoder
H_r
Relation Description Generation
c) Contrastive Semantic Retrieval Strategy

	H_r^1	H_r^2	H_r^3	H_r^4	H_r^5	H_r^6
H_e^1	1	0	0	0	0	0
H_e^2	0	1	0	0	0	0
H_e^3	0	0	1	0	0	0
H_e^4	0	0	0	1	0	0
H_e^5	0	0	0	0	1	0
H_e^6	0	0	0	0	0	1

Figure 2: A multimodal relational extraction model based on relational semantic retrieval paradigm.

visual features X_I. Then, we concatenate the textual and visual features along the sequence dimension to form a unified multimodal representation $X = [X_T; X_I]$, which is passed through multiple Transformer encoder layers. A multi-head self-attention mechanism is applied to achieve deep cross-modal semantic fusion:

$$H_X = \text{Encoder}(XW_Q, XW_K, XW_V)_N \tag{1}$$

Where W_Q, W_K, W_V are the learnable projection matrices for queries, keys, and values in the self-attention mechanism.

Let the indices of the subject and object entities in the concatenated sequence be $\tilde{s}$ and $\tilde{o}$. Based on their positions in the text, we extract the contextualized representations (H_s, H_o) of the subject entity s and the object entity o from H_X:

$$\begin{aligned} H_s &= H_X[\tilde{s}], H_s \in \mathbb{R}^H \\ H_o &= H_X[\tilde{o}], H_o \in \mathbb{R}^H \end{aligned} \tag{2}$$

Finally, the entity pair representation is fused via a fully connected layer with a nonlinear activation to obtain the multimodal entity representation used for relation prediction:

$$H_E = \sigma(W_e[H_s; H_o] + b_e) \tag{3}$$

where $H_E \in \mathbb{R}^H$ is entity-pair representation, σ is non-linear activation function, and $W_e \in \mathbb{R}^{H \times 2H}$ and $b_e \in \mathbb{R}^H$ are learnable parameters.

3.4 Relational Semantic Encoder

To enhance the model's ability capturing relational semantics, we introduce natural language descriptions to replace traditional discrete label representations which enables the model to precisely distinguish semantic differences among relation types.

We first utilize the GPT-4o model to convert each relation label in the training dataset into a natural language description. The generated descriptions are then manually reviewed to ensure both accuracy and semantic consistency.

For each relation description d_i, we use an independent BERT encoder (denoted as `RelEncoder`) to encode the description and obtain a global semantic representation of the relation:

$$\begin{aligned} X_r &= \text{RelEncoder}(d_i) \\ H_r &= \frac{1}{L_R} \sum_{i=1}^{L_R} X_r[i] \end{aligned} \tag{4}$$

Where H_r denotes the mean-pooled semantic vector representing the relation, which captures the distribution of the relation in the semantic space.

3.5 Contrastive Semantic Retrieval Strategy

During model training, the ROC framework abandons traditional classification loss functions and instead constructs a contrastive learning-based semantic retrieval mechanism. Inspired by the SimCLR approach, we optimize the model's ability to discriminate semantic relations by maximizing the cosine similarity between positive samples (i.e., an entity pair and its corresponding relation description) while minimizing the similarity to negative relation descriptions within the same batch.

The loss function is defined as follows:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathrm{sim}(H_E^i, H_r^i)/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(H_E^i, H_r^j)/\tau)} \tag{5}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is a temperature hyperparameter, and H_E^i and H_r^i represent the multimodal entity pair representation and the corresponding relation semantic representation for the i-th sample, respectively.

3.6 Inference

During inference phase, the model calculates the similarity between the multimodal entity pair representation and all relation semantic embeddings, and selects the relation with the highest similarity as the final prediction. This retrieval-based inference approach significantly enhances the interpretability and flexibility of relation extraction, allowing the model to perform dynamic matching based on semantic similarity rather than relying on fixed category labels. Compared to traditional classification methods, this approach demonstrates stronger adaptability and generalization when handling relation semantic shifts or unseen relation types.

4 Experimental Settings

Dataset We evaluate our method on two widely used multimodal relation extraction datasets: MNRE(Zheng et al., 2021b) and MORE(He et al., 2023), which respectively correspond to social media contexts and object-entity relation extraction tasks. These datasets encompass rich information on image-text alignment and diverse relation types. More details can be found in Appendix B.

Evaluation Metrics We adopt four commonly used evaluation metrics for multimodal relation extraction tasks: Accuracy, Precision, Recall, and F1 score. The F1 score, which balances Precision and Recall, is the main criterion for subsequent performance analysis.

Baselines We compare ROC with a range of representative multimodal RE baselines. Early methods include MTB(Soares et al., 2019b), VisualBERT(Li et al., 2019a), ViLBERT(Lu et al., 2019), UMT(Yu et al., 2020c), MKGformer(Chen et al., 2022a), MEGA(Zheng et al., 2021a) and HVPNeT(Chen et al., 2022b). Among more recent advances, IFAformer(Li et al., 2023c) improves visual-textual alignment through prefix networks and early cross-attention. TSVFN(Zhao et al., 2023) employs a two-stage fusion strategy to mitigate visual noise. PROMU(Hu et al., 2023) and MOREformer(He et al., 2023) enhance relation prediction via prompt-based and object-centric designs. TMR(Zheng et al., 2023) leverages diffusion-based generation for robust alignment, while MMIB(Cui et al., 2024) adopts an information bottleneck to reduce modality noise. VM-HAN(Li et al., 2024) models higher-order relations using multimodal hypergraphs, and CAMRE(Zhang et al., 2024) introduces LLM-generated image descriptions to improve alignment. APOLLO(Zhang et al., 2025) proposes a triple contrastive mechanism for cross-modal semantic learning. We also use Qwen-VL(Bai et al., 2025), BLIP2(Li et al., 2023b), and InstructBLIP(Dai et al., 2023) as vision-language LLM baselines to further validate ROC's effectiveness.

Implementation details For detailed information on model configuration, training setup, and hyperparameter settings, see Appendix C.

5 Experimental Results

5.1 Main Results

To comprehensively evaluate the performance of our proposed ROC model on the multimodal relation extraction task, we conducted main experiments on two standard datasets: MNRE and MORE, and compared our method with several representative existing models. The experimental results are shown in Table 1 and Table 2.

On the MNRE dataset, the ROC model achieved an accuracy of 90.97%, which is lower than CAMRE (95.79%). However, it outperformed all other methods in terms of recall (90.85%) and F1 score (91.22%). Compared with CAMRE, ROC improved recall by 0.69 percentage points (a relative improvement of 0.77%) and F1 score by 0.28 percentage points (a relative improvement of 0.31%). Significance testing on the F1 score shows a 95% confidence interval of [90.93%, 91.51%], indicating that the improvement is statistically significant. Since both models already exceed 90% on key metrics, even minor improvements demonstrate ROC's stronger ability in identifying positive samples in multimodal relation extraction.

It is worth noting that models such as TMR and CAMRE utilize additional information (e.g., synthetic samples or image descriptions generated by large models) to enhance understanding of image content, thereby improving the accuracy of relation prediction between entities. However, these methods often overlook the modeling of negative

Model	Accuracy	Precision	Recall	F1-score
MTB (Soares et al., 2019b)	75.69	64.46	57.81	60.86
UMT (Yu et al., 2020c)	77.84	62.93	63.88	63.46
MEGA (Zheng et al., 2021a)	80.05	64.51	68.44	66.41
HVPNeT (Chen et al., 2022b)	92.52	83.64	80.78	81.25
IFAformer (Li et al., 2023c)	92.38	82.59	80.78	81.67
TSVFN (Zhao et al., 2023)	92.67	85.16	82.07	83.12
MOREformer (He et al., 2023)	82.67	82.19	82.35	82.27
PROMU (Hu et al., 2023)	–	84.95	85.76	84.86
TMR (Zheng et al., 2023)	–	90.48	87.66	89.05
MMIB (Cui et al., 2024)	–	83.49	82.97	83.23
VM-HAN (Li et al., 2024)	92.57	85.76	84.69	85.22
CAMRE (Zhang et al., 2024)	**95.79**	**91.73**	90.16	90.94
ROC (Ours)	90.97 (±0.32)	91.59 (±0.24)	**90.85** (±0.32)	**91.22** (±0.23)

Table 1: Main experimental results of the ROC model on the MNRE dataset.

Model	Accuracy	Precision	Recall	F1-score
BERT+SG	61.79	29.61	41.27	34.48
BERT+SG+Att	63.74	31.10	39.28	34.71
MEGA (Zheng et al., 2021a)	65.97	33.30	38.53	35.72
IFAformer (Li et al., 2023c)	79.28	55.13	54.24	54.68
MKGformer (Chen et al., 2022a)	80.17	55.76	53.74	54.73
VisualBERT (Li et al., 2019a)	82.84	58.18	61.22	59.66
ViLBERT (Lu et al., 2019)	83.50	62.53	59.73	61.10
MOREformer (He et al., 2023)	83.50	62.18	63.34	62.75
VM-HAN (Li et al., 2024)	85.57	64.76	66.69	65.71
APOLLO (Zhang et al., 2025)	85.90	67.42	70.70	69.02
ROC (Ours)	**90.44** (±0.31)	**68.85** (±1.12)	**75.40** (±0.57)	**71.97** (±0.86)

Table 2: Main experimental results of the ROC model on the MORE dataset.

samples (i.e., the "None" relation), resulting in limited improvements in recall. The significant recall improvement achieved by ROC indicates that the retrieval-based paradigm incorporating relation semantics allows for a deeper understanding of semantic relations between entities, leading to more accurate predictions than traditional classification approaches relying solely on discrete labels.

The ROC model outperformed existing methods across all evaluation metrics on the more challenging MORE dataset, demonstrating state-of-the-art performance in relation prediction. Specifically, ROC achieved an F1 score of 71.97, outperforming the second-best model APOLLO by 2.95 percentage points (a relative improvement of 4.27%). In terms of recall, it reached 75.40, an improvement of 4.70 percentage points over APOLLO (a relative improvement of 6.65%), while accuracy and precision also increased by 5.29% and 2.12%, respectively. These results confirm that the retrieval-based paradigm with relation semantics helps the model more comprehensively and accurately predict semantic relations between entities.

Overall, the retrieval-based multimodal relation extraction approach employed by the ROC model effectively aligns entity pairs with their potential semantic relationships. It achieves the best overall F1 scores on both the MNRE and MORE datasets, providing strong evidence of the effectiveness of the ROC model design.

5.2 Ablation Study

To evaluate the contribution of each core component in the ROC model to the overall performance, we designed four ablation studies by removing key model modules and observing the resulting performance changes. The experimental results are shown in Table 3.

- w/o entity encoder: Removes the Transformer encoder in the multimodal entity pair feature encoding module to assess the effect of cross-modal feature interaction and fusion.

- w/o entity position: Removes the entity position encoding mechanism to evaluate the impact of positional information on relation prediction. Global average pooling is applied to maintain feature dimensional consistency.

Ablation Setting	MNRE				MORE			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
ROC (Full Model)	90.97	91.59	90.85	91.22	90.44	68.85	75.40	71.97
w/o entity encoder	90.40	90.49	90.74	90.62	87.09	62.42	73.94	67.69
Δ	-0.57	-1.1	-0.11	-0.6	-3.35	**-6.43**	-1.46	-4.28
w/o entity position	90.15	91.00	90.12	90.56	89.10	66.22	73.32	69.59
Δ	-0.82	-0.59	-0.73	-0.66	-1.34	-2.63	-2.08	-2.38
w/o entity type	89.16	89.57	88.96	89.26	88.37	65.60	71.82	68.57
Δ	-1.81	-2.02	-1.89	-1.96	-2.07	-3.25	-3.58	-3.4
w/o relation embedding	88.23	89.00	87.72	88.36	84.85	63.88	64.84	64.36
Δ	**-2.74**	**-2.59**	**-3.13**	**-2.86**	**-5.59**	-4.97	**-10.56**	**-7.61**

Table 3: Ablation results of the ROC model on MNRE and MORE datasets. Each row removes one component from the full model. Δ indicates the performance drop compared to the full model.

- w/o entity type: Removes the pre-extracted entity type information to assess the influence of type priors on prediction performance.
- w/o relation embedding: Replaces the relation semantic encoder with fixed relation label IDs, degrading the model into a conventional classification architecture to evaluate the effectiveness of the semantic retrieval paradigm.

According to the results, removing the cross-modal interaction layer (w/o enc) decreased precision by 1.10 and 6.43 percentage points on the MNRE and MORE datasets, respectively. This indicates that the lack of explicit cross-modal feature interaction significantly degrades performance.

Removing the entity position encoding (w/o ent-pos) caused a performance drop across all metrics, with recall declining even more than in w/o enc. This suggests that positional encoding plays a critical role in relation prediction. On one hand, it explicitly marks the positions of the subject and object entities, enhancing the model's ability to distinguish entity pair structures. On the other hand, it constrains the semantic scope for relation modeling, preventing indiscriminate matching based on global text and image features. Without it, the model must rely on global cues to retrieve relations, lowering prediction accuracy and coverage.

When the entity type embedding was removed (w/o ent-type), all metrics on the MNRE dataset declined, with a 2.02 percentage point drop in precision—the second largest drop among all ablation settings. Precision and recall on the MORE dataset decreased by 3.25 and 3.58 percentage points, respectively, with recall experiencing the second-largest decline. These results indicate that entity type information effectively constrains the relational semantic space. Since relation semantics involve explicit meaning and imply subject-object roles and type expectations, entity types help form a constraint-verification mechanism with relation semantics. Without this information, the model struggles to distinguish between semantically similar relations but differ in kind, which harms prediction accuracy.

After removing the relation semantic encoder (w/o relation), the model experienced an average drop of nearly 3 percentage points across all metrics on the MNRE dataset. On the MORE dataset, recall fell by 10.56 percentage points and F1 score by 7.61 points, indicating even more substantial performance degradation. This powerfully demonstrates the critical role of explicit relation semantic modeling in improving multimodal relation extraction accuracy. The relation semantic encoder provides fine-grained semantic constraints, enabling the model to perform relation prediction via semantic matching. Without this module, the model relies solely on fused features for classification, lacking clear semantic guidance. This increases decision uncertainty, especially in cross-modal subject-object scenarios like the MORE dataset.

Overall, the submodules in the ROC model work collaboratively to build a robust cross-modal semantic matching mechanism.

5.3 Effect of Visual Input on ROC Performance

To evaluate the impact of visual information on model performance, we conducted a controlled experiment by removing the visual modality. The results in Figure 3 show that the model's performance on the MNRE dataset remains largely unaffected after excluding image inputs. In contrast, a significant performance drop is observed on the MORE dataset. This suggests that the MORE dataset relies

Method	MNRE				MORE			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Retrieval-based	90.97	91.59	90.85	91.22	90.44	68.85	75.40	71.97
Classification-based	74.91	76.99	75.51	76.25	74.24	45.41	51.87	48.43

Table 4: Comparison of retrieval-based and classification-based methods on MNRE and MORE datasets.

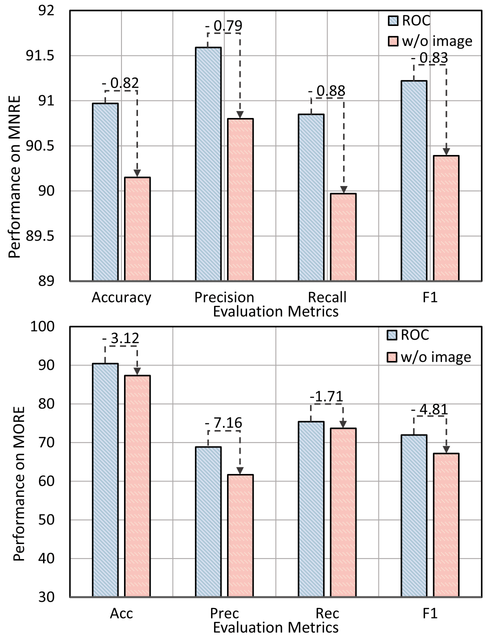


Figure 3: Impact of visual information on ROC performance on MNRE and MORE datasets.

more heavily on visual information and further validates the effectiveness of deep multimodal fusion in high-quality vision-language matching tasks.

5.4 Comparison of Retrieval-based and Classification-based Methods

To further validate the retrieval-based approach over the classification-based one, we replaced the retrieval module with a standard classification head (a fully connected layer followed by softmax) to compare the performance of both versions. Results are shown in Table 4.

As shown in the results, the classification-based version exhibits a significant drop across all metrics. On the MNRE and MORE datasets, the F1 scores decrease by 14.97 and 23.54 percentage points, respectively. This performance gap is primarily because the retrieval-based method leverages semantic matching between the relation descriptions and the input context, while also using entity types to effectively narrow down the candidate space. In contrast, the classification model must select from all possible relation types and cannot fully exploit such contextual information.

5.5 Comparison with MLLMs

To further evaluate the advantages of the ROC model under the current trend of large multimodal language models (MLLMs), we compare it with several representative MLLMs, including fine-tuned versions of BLIP2, InstructBLIP, and Qwen-VL-Plus, as well as the non-fine-tuned DeepSeek-V3. The comparison results are shown in Table 5.

On the MNRE dataset, the fine-tuned MLLMs achieve high precision, with Qwen-VL-Plus reaching 95.57%, and InstructBLIP and BLIP2 achieving 94.98% and 94.86%, respectively. However, regarding F1 score, the ROC model outperforms all MLLMs, indicating a more balanced performance between precision and recall. Although MLLMs perform competitively in some metrics, they still lag behind ROC in recall and overall stability.

On the more challenging MORE dataset, MLLMs' accuracy remains relatively high, but both precision and recall drop significantly, leading to much lower F1 scores compared to ROC. This performance gap may be attributed to increased semantic complexity, which makes MLLMs more prone to overfitting on the training set and less capable of generalizing to diverse samples.

DeepSeek-V3, as non-fine-tuned MLLM, only performs zero-shot inference in experiments. Its performance is significantly worse than the fine-tuned models and ROC, suggesting that current MLLMs struggle to handle structured extraction tasks without task-specific adaptation.

While MLLMs can achieve competitive results under sufficient resources and tuning conditions, their training costs and adaptation thresholds are relatively high. In contrast, with its lightweight design, the ROC model achieves the best overall performance on both datasets and consistently leads in key metrics such as F1 score, demonstrating superior practicality and deployability.

Additional analyses, including prompt templates, encoder architecture variations, and attention distributions, are provided in Appendix E–I.

Model	MNRE				MORE			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
BLIP2 (Li et al., 2023b)	94.86	89.42	89.84	89.63	73.09	29.91	35.54	32.48
InstructBLIP (Dai et al., 2023)	94.98	90.11	89.69	89.90	79.48	49.33	55.24	52.12
QwenVL-Plus (Bai et al., 2025)	**95.57**	91.41	89.84	90.62	-	-	-	-
DeepSeek-V3* (Liu et al., 2024a)	21.62	21.65	54.53	30.99	16.70	16.73	41.13	23.79
ROC (Ours)	90.97	**91.59**	**90.85**	**91.22**	**90.44**	**68.85**	**75.40**	**71.97**

Table 5: Comparison of the ROC model with MLLMs on the MNRE and MORE datasets.

6 Conclusion

We introduce a relation-semantic retrieval-based method for multimodal relation extraction, named ROC. It integrates entity-centric multimodal encoding, position-aware structural modeling, and relation-aware semantic retrieval, showing robust performance across diverse scenarios. Experiments demonstrate that ROC outperforms baselines on the MNRE and MORE datasets, including fine-tuned large-scale MLLMs, particularly excelling in F1 scores. Ablation studies further confirm the contributions of each key component, highlighting the critical roles of explicit multimodal interaction and structured semantic modeling. ROC overcomes the limitations of traditional classification methods in label semantic representation and fine-grained semantic differentiation, offering a novel paradigm for multimodal relation extraction.

Limitations

The limitations of our approach are as follows: Our experiments show that zero-shot multimodal LLMs cannot directly perform multimodal relation extraction. However, we have not yet systematically verified whether fine-tuned LLMs can significantly enhance task performance, especially when class labels are reformulated as natural language descriptions. Furthermore, although replacing discrete class labels with semantic descriptions is theoretically applicable to a wide range of classification tasks, the generalizability of this method has not been thoroughly evaluated across diverse domains, tasks, and datasets.

Ethics Statements

Our model infers potential relations between entities from text and images, but these are based solely on input content and do not reflect verified real-world facts. The datasets may include personal information and perform basic checks for identifiable or offensive content, but named entities central to the task cannot be anonymized. No human annotators or evaluators are involved; all experiments and evaluations are automated. GPT-4o generates the relation descriptions. The AI tools are used only for grammar correction and relation description generation.

Acknowledgements

This work is funded by the National Natural Science Foundation of China (No.62576085), Zhiyuan Laboratory (NO. ZYL2024021) and CCF-Zhipu AL Large Model Fund (NO.202221).

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. In *International Conference on Learning Representations*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022a. 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.

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022b. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1607–1618.

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

Shiyao Cui, Jiangxia Cao, Xin Cong, Jiawei Sheng, Quangang Li, Tingwen Liu, and Jinqiao Shi. 2024. Enhancing multimodal entity and relation extraction with variational information bottleneck. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1274–1285.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Liang He, Hongke Wang, Yongchang Cao, Zhen Wu, Jianbing Zhang, and Xinyu Dai. 2023. More: A multimodal object-entity relation extraction dataset with a benchmark evaluation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4564–4573.

Xuming Hu, Junzhe Chen, Aiwei Liu, Shiao Meng, Lijie Wen, and Philip S Yu. 2023. Prompt me up: Unleashing the power of alignments for multimodal entity and relation extraction. In *Proceedings of the 31st ACM international conference on multimedia*, pages 5185–5194.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bo Li, Dingyao Yu, Wei Ye, Jinglei Zhang, and Shikun Zhang. 2023a. Sequence generation with label augmentation for relation extraction. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Lei Li, Xiang Chen, Shuofei Qiao, Feiyu Xiong, Huajun Chen, and Ningyu Zhang. 2023c. On analyzing the role of image for visual-enhanced relation extraction (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16254–16255.

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

Qian Li, Cheng Ji, Shu Guo, Yong Zhao, Qianren Mao, Shangguang Wang, Yuntao Wei, and Jianxin Li. 2024. Variational multi-modal hypergraph attention network for multi-modal relation extraction. *ACM Multimedia 2024*.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019b. Entity-relation extraction as multi-turn question answering. *arXiv preprint arXiv:1905.05529*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Xiyang Liu, Chunming Hu, Richong Zhang, Kai Sun, Samuel Mensah, and Yongyi Mao. 2024b. Multimodal relation extraction via a mixture of hierarchical visual context learners. In *Proceedings of the ACM Web Conference 2024*, pages 4283–4294.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Kangqi Luo, Fengli Lin, Xusheng Luo, and Kenny Zhu. 2018. Knowledge base question answering via encoding of complex query graphs. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2185–2194.

Jinzhong Ning, Zhihao Yang, Yuanyuan Sun, Zhizheng Wang, and Hongfei Lin. 2023. OD-RTE: A one-stage object detection framework for relational triple extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11120–11135, Toronto, Canada. Association for Computational Linguistics.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, and ... Adam Perelman. 2024. Gpt-4o system card. *arXiv preprint*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020a. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020b. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Feiliang Ren, Longhui Zhang, Shujuan Yin, Xiaofeng Zhao, Shilei Liu, Bochao Li, and Yaduo Liu. 2021a. A novel global feature-oriented relational triple extraction model based on table filling. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2646–2656, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Feiliang Ren, Longhui Zhang, Xiaofeng Zhao, Shujuan Yin, Shilei Liu, and Bochao Li. 2021b. A simple but effective bidirectional extraction framework for relational triple extraction. *CoRR*, abs/2112.04940.

Yu-Ming Shang, Heyan Huang, Xin Sun, Wei Wei, and Xian-Ling Mao. 2022. Relational triple extraction: One step is enough. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4360–4366. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

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

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

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725.

Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. 2023. 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)*, pages 14734–14751.

Bowen Yu, Xue Mengge, Zhenyu Zhang, Tingwen Liu, Wang Yubin, and Bin Wang. 2020a. Learning to prune dependency trees with rethinking for neural relation extraction. In *Proceedings of the 28th international conference on computational linguistics*, pages 3842–3852.

Haoze Yu, Haisheng Li, Dianhui Mao, and Qiang Cai. 2020b. A relationship extraction method for domain knowledge graph construction. *World Wide Web*, 23(2):735–753.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020c. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352.

Li Yuan, Yi Cai, Jin Wang, and Qing Li. 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, pages 11051–11059.

Yaming Zhang, Jianfei Yu, Wenya Wang, Li Yang, Jia Yang, and Rui Xia. 2025. Enhancing multimodal object-entity relation extraction via multi-aspect contrastive learning in large multimodal models. *IEEE Transactions on Audio, Speech and Language Processing*, 33:1220–1229.

Zefan Zhang, Weiqi Zhang, Yanhui Li, and Tian Bai. 2024. Caption-aware multimodal relation extraction with mutual information maximization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1148–1157.

Qihui Zhao, Tianhan Gao, and Nan Guo. 2023. Tsvfn: Two-stage visual fusion network for multimodal relation extraction. *Information Processing & Management*, 60(3):103264.

Changmeng Zheng, Junhao Feng, Yi Cai, Xiaoyong Wei, and Qing Li. 2023. 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)*, pages 6810–6824.

Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021a. Multimodal relation extraction with efficient graph alignment. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5298–5306.

Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 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)*, pages 1–6. IEEE.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168, Online only. Association for Computational Linguistics.

A Licenses

The models used in this work, including ViT-B/16 (Dosovitskiy et al., 2021) and BERT (Devlin et al., 2019), are licensed under the Apache License 2.0. The GPT series models are developed and released by OpenAI under their respective terms of use. Detailed license information is available on the official GitHub repositories or documentation pages.

We use the Stanford Named Entity Recognizer (NER) tool for entity recognition, which is distributed under the GNU General Public License v2 (GPLv2).[1]

The datasets used include the MNRE dataset(Zheng et al., 2021b), with details available on its GitHub page, and the MORE dataset(He et al., 2023), which is released under the MIT License.

In summary, all licenses permit academic research use.

B Dataset Details

In the context of increasingly rich multimedia data, extracting structured information from multimodal data that includes both text and images has become particularly important. To promote research in this field, the academic community has introduced specially designed datasets to support the development of relation extraction tasks. **MNRE** (Zheng et al., 2021b) is a specially designed dataset aimed at evaluating and enhancing the capabilities of neural relation extraction models, with a particular emphasis on the importance of incorporating visual evidence in social media posts. The dataset contains over 9,000 sentences covering 23 distinct relation types, sourced from Twitter and annotated by crowd-sourced workers. Each sentence is paired with a relevant image, intended to supplement contextual information that may be missing from the text alone, thereby aiding in the more accurate identification of relationships between entities. **MORE** (He et al., 2023) is a novel dataset focused on extracting object-entity relations from both text and images, developed by a research team from Nanjing University. It consists of 3,559 pairs of news headlines and their corresponding images, annotated with 20,264 multimodal relational facts across 21 relation types, involving 13,520 visual objects with an average of 3.8 objects per image. MORE is designed to pose challenges to existing methods in handling complex relationships between text and images, particularly emphasizing scenarios that require identifying relations between entities and visual objects across different modalities. This dataset serves as an important resource for advancing research on multimodal relation extraction. With MORE, researchers can explore how to enhance models' ability to understand the interactions between textual and visual information.

C Implementation Details

To ensure the fairness and rigor of our conclusions, we adopted the same text encoder as used in previous methods. We conducted comprehensive experiments on both the MNRE and MORE datasets. Specifically, the model was trained for 50 epochs with a batch size of 32, using the AdamW optimizer and a hidden layer dimension 768. The overall model contains approximately 342.52 million parameters, including the BERT-base-uncased text encoder (109.48 million parameters), the ViT-base-patch32-384 image encoder (88.12 million parameters), and the BERT-base-uncased relation encoder (109.48 million parameters). To ensure reproducibility, all experiments were conducted under the following setup: CPU was Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz with 32 cores, memory size was 128GB, and GPU was NVIDIA RTX 8000 with 48GB VRAM. The operating system was Ubuntu 16.04.7 LTS, with CUDA version 11.7, PyTorch version 1.13.1, and Python version 3.10.4.

The experimental results are obtained by one-shot inference with a random seed of 648, and the results are reproducible.

[1] `https://nlp.stanford.edu/software/CRF-NER.html`

Dataset	#Image	#Word	#Sentence	#Entity	#Relation	#Instance
MNRE	9,201	258k	9,201	30,970	23	15,485
MORE	3,559	-	3,559	-	21	3,559

Table 6: Detailed information on MNRE and MORE datasets.

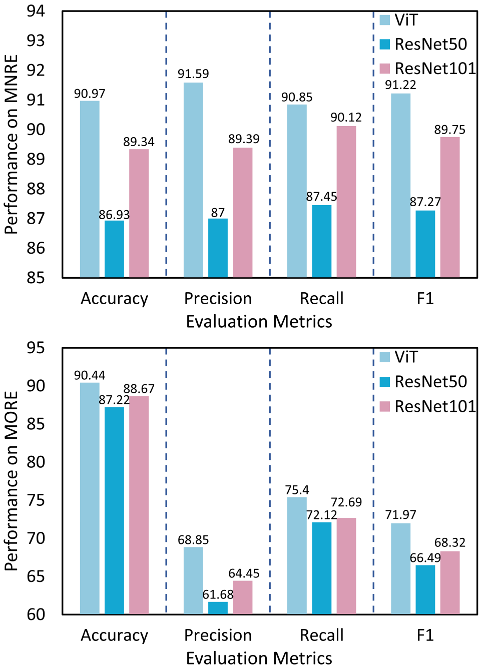


Figure 4: Effect of visual encoder on relation prediction performance on MNRE and MORE datasets.

D Prompt Template for Extending Natural Language Descriptions

This study employs GPT-4o to generate explanations and descriptions for the predefined set of relations in the MNRE and MORE datasets. Table 7 presents the specific prompt and several example relation descriptions generated by large model.

E Impact of Visual Encoder Architectures on Model Performance

To evaluate the performance of different visual encoders in multimodal relation extraction tasks, we replaced the visual feature extraction module in the ROC model with ResNet50, ResNet101 (CNN-based architectures), and ViT-B/16 (a Transformer-based architecture). We conducted comparative experiments on the MNRE and MORE datasets, and the results are shown in Figure 4.

On the MNRE dataset, ViT outperforms the ResNet-based encoders across all metrics. Specifically, its F1 score is 3.95 percentage points higher than that of ResNet50 and 1.47 points higher than that of ResNet101, indicating stronger stability and generalization capability. On the more complex MORE dataset, which has a higher dependency on visual semantics, ViT achieves even greater improvements, with F1 scores 5.48 and 3.65 percentage points higher than ResNet50 and ResNet101, respectively. The precision gain is particularly notable, with a 7.17-point increase over ResNet50.

The primary reason for this performance difference lies in the architectural consistency and its impact on modality fusion. In the ROC model, the text encoder and the relation semantics encoder are based on the Transformer architecture. ViT, as a structurally homogeneous visual encoder, adopts a similar approach to feature extraction and semantic modeling as BERT, utilizing self-attention mechanisms to model global dependencies. This architectural compatibility naturally facilitates more efficient semantic alignment in the fusion stage.

In contrast, as CNN-based encoders, the ResNet series emphasize local feature extraction. Their localized modeling mechanisms are structurally heterogeneous to the global modeling strategy of Transformers, introducing additional alignment challenges during fusion and thus reducing the quality of modality integration. Since multimodal relation extraction tasks rely heavily on global semantic reasoning between text and image, Transformer-based encoders are inherently more suitable.

F Impact of Encoder Depth on Model Performance

As a key component of modality fusion, the Transformer encoder in the ROC model models the attention-based interactions between visual and textual information. This section investigates the impact of different encoder depth on model performance through comparative experiments, with the results shown in Figure 5.

When no encoder is introduced on the MNRE dataset, the model achieves an accuracy of 90.40% and an F1 score of 90.62%. With the addition of a single encoder layer, these metrics increase to 90.52% and 90.81%, respectively, indicating that the attention mechanism contributes positively to

Relation Description Few-shot Prompt Template

Given a set of relation labels, convert them into a JSON format where each key is a relation label and the value is a clear, textual explanation of the relationship that the label represents. If the provided labels include numeric indices, please remove them in your output.
There are some relation description examples:
{
"None": "Indicates that there is no relationship between the subject and object entity, based on text and image information. Subject and object can be of any type.",
"/org/loc/locate_at": "Indicates that an organization is located at a specific geographical location (e.g., a company headquartered in a city)",
"/per/misc/awarded": "Specifies an award, honor, or recognition received by a person (e.g., a Nobel Prize or medal)."
}

Example of Generated Relation Descriptions and Label Set

{
"None": "Indicates that there is no relationship between the subject and object entity, based on text and image information. Subject and object can be of any type.",
"/loc/loc/contain": "Describes a containment relationship where both the subject and object are geographical locations, with the subject containing the object (e.g., a country containing a city).",
"/misc/loc/held_on": "Indicates that a miscellaneous event or activity occurred at a specific geographical location (e.g., a festival held in a city).",
"/misc/misc/part_of": "Describes a part-whole relationship between two miscellaneous entities (e.g., a chapter being part of a book).",
"/org/loc/locate_at": "Indicates that an organization is located at a specific geographical location (e.g., a company headquartered in a city).",
"/org/org/alternate_names": "Lists alternative names, aliases, or abbreviations for an organization (e.g., a company and its trade name).",
"/org/org/subsidiary": "Describes a parent-subsidiary relationship between two organizations, with the subject being the parent and the object being the subsidiary (e.g., a corporation and its owned companies).",
"/per/loc/place_of_birth": "Indicates the geographical location where a person was born (e.g., a city or country)."
}

Table 7: Prompt template and generated relation descriptions for few-shot relation label explanation.

SciRE		**NYT**	
Model	F1	Model	F1
MTB (Soares et al., 2019b)	87.4	BiRTE (Ren et al., 2021b)	92.8
REBEL (Huguet Cabot and Navigli, 2021)	87.7	DirectRel (Shang et al., 2022)	92.9
IRE-RoBERTa (Zhou and Chen, 2022)	88.9	GRTR (Ren et al., 2021a)	93.4
RELA (Li et al., 2023a)	90.3	OD-RTE (Ning et al., 2023)	93.9
ROC (ours)	88.39	ROC (ours)	91.36

Table 8: Performance comparison on the SciRE and NYT datasets. SciRE results are from (Li et al., 2023a), and NYT results are from (Ning et al., 2023)

basic multimodal fusion. The model performance improves as the encoder depth increases to 3 and 6 layers. However, the F1 score of the 3-layer configuration slightly surpasses that of the 6-layer one, with an improvement of 0.41 percentage points. This difference is primarily due to a slight drop in recall for the 6-layer structure, suggesting that while deeper encoders enhance prediction precision, they may also suppress the recognition of marginal samples, thereby affecting recall.

On the MORE dataset, the model shows a more significant response to increasing encoder depth. Without any encoder, the model achieves an accuracy of 87.09% and an F1 score of 67.69%. When the encoder depth is increased to 6 layers, the F1 score improves by 4.55 percentage points, with all metrics reaching their highest values. This indicates that deep attention mechanisms play a substantial role in enhancing cross-modal semantic fusion on this dataset.

The performance differences between the two datasets can be attributed mainly to the varying quality of image-text alignment. The MNRE dataset suffers from relatively weak semantic associations between images and text, where shallow fusion helps preserve more original semantics and allows the model to learn alignment strategies autonomously. In contrast, the MORE dataset is constructed from news articles with well-aligned image-text pairs, where deeper multimodal interaction more effectively captures applicable semantics. Therefore, deeper encoder layers are more benefi-

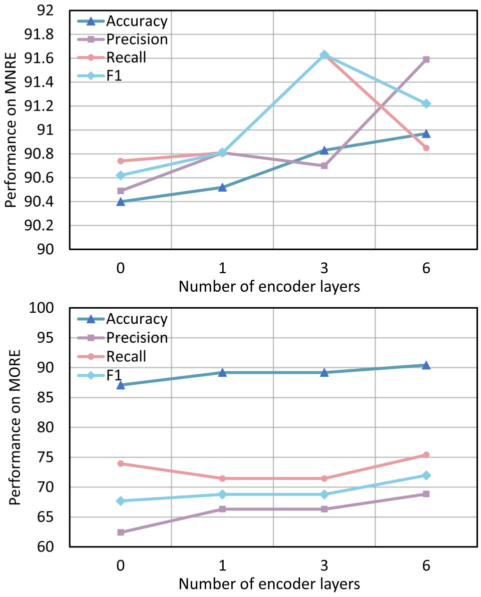


Figure 5: Impact of different encoder numbers on modal fusion performance on MNRE and MORE datasets.

cial for modeling complex semantic relationships in multimodal scenarios with structurally complete and semantically consistent inputs.

In summary, the optimal depth of the encoder should be adapted to the dataset's characteristics: shallow structures are better suited for tasks with loose image-text alignment, while deeper structures are more effective in scenarios with strong semantic coupling between modalities, promoting semantic aggregation and relation understanding.

G Experiments on Text-Only Datasets

We conducted experiments on two text-only datasets from different domains: SciRE, which focuses on scientific papers, and NYT, which covers the news domain, in order to evaluate the applicability of ROC across diverse professional text datasets. The results are presented in Table 8.

As shown in the results, on the SciRE dataset, ROC performs only 1.54 percentage points lower than RELA, while outperforming MTB and REBEL. On the NYT dataset, ROC is 2.54 percentage points lower than OD-RTE. These findings demonstrate that ROC remains competitive and stable across different domain-specific text datasets.

H Case Study

To evaluate the ability of the ROC model to distinguish semantically similar relations, we selected a subset of closely related relations from the MNRE and MORE datasets and constructed relation pairs.

MNRE	**Accuracy**
/per/misc/nationality vs /per/misc/race	100.00
/per/per/peer vs /per/per/neighbor	94.94
/org/loc/locate_at vs /loc/loc/contain	92.41
/per/misc/present_in vs /per/loc/place_of_residence	90.29
/per/org/member_of vs /per/per/alumi	90.00
MORE	**Accuracy**
/org/loc/locate_at vs /org/misc/present_in	100.00
/per/misc/president vs /per/org/leader_of	97.73
/per/misc/nationality vs /per/misc/party	98.97
/per/per/relatives vs /per/per/partner	92.65
/per/misc/present_in vs /org/misc/present_in	91.74

Table 9: Accuracy of ROC on semantically related relation pairs from the MNRE and MORE datasets.

In the experiments, the model was required to identify the correct relation within each pair, effectively performing a binary classification task. The experimental results are shown in Table 9.

As shown in the results, ROC attains consistently high accuracy across most relation pairs. For pairs with clear semantic distinctions, such as `/per/misc/nationality-/per/misc/race` and `/org/loc/locate_at-/org/misc/present_in`, the model achieves 100% accuracy. For pairs that are semantically similar but differ in entity types, such as `/per/misc/present_in-/per/loc/place_of_residence` and `/per/misc/present_in-/org/misc/present_in`, the accuracies are 90.29% and 91.74%, respectively, demonstrating that ROC can effectively exploit entity type information to distinguish subtle relation differences.

I Visualization of Attention Weight Distribution

To further validate the ROC model's semantic modeling capability and interpretability in processing specific samples, we conducted a visualization analysis of the attention distribution in the feature encoder for multimodal entity pairs. As shown in Figure 6, a real-world example was selected: *"RT @DenisLlaw_WFT: New breed of Crocodile discovered in South Wales woodland",* where the subject is "Crocodile", the object is "South Wales", and the relation type is `/misc/loc/held_on`.

The figure illustrates the attention weight distribution of query vectors across different encoding layers. In layers 1 and 2, the attention distribution is relatively dispersed and does not focus on key entities, indicating that at this stage, the model mainly captures global semantic features without explicitly

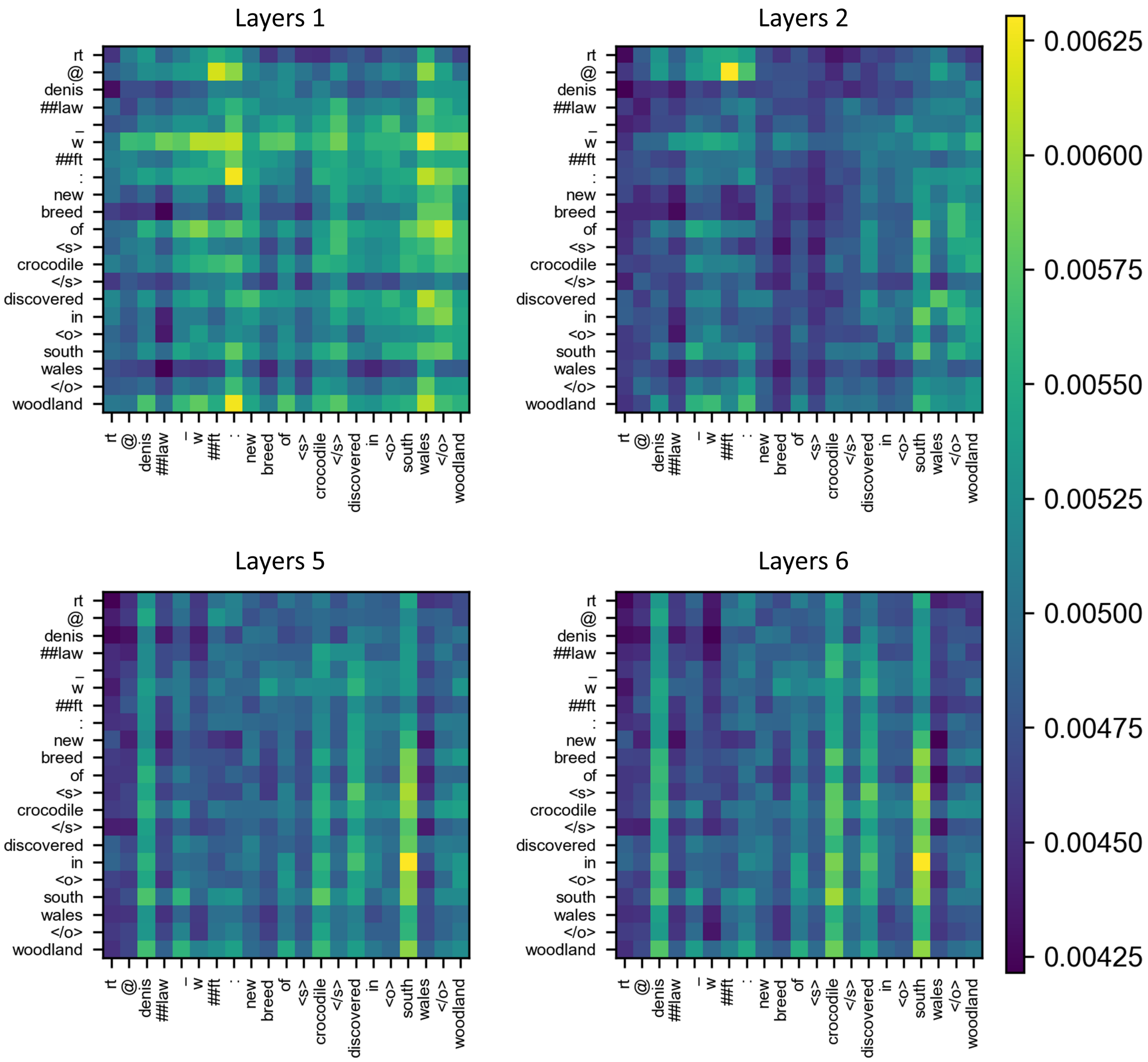


Figure 6: Visualization of attention weight distribution.

concentrating on the entity pair. In contrast, layers 5 and 6 show a progressively focused attention pattern, with most attention weights concentrated on the words "crocodile" and "south", which correspond to the start positions of the entity pair. This demonstrates the model's dynamic transition from integrating global semantics to identifying local entities layer by layer.

Moreover, in layers 5 and 6, the query word "in" exhibits high attention towards the object word "south", consistent with the semantic alignment of "loc" in the predicted relation `/misc/loc/held_on`, indicating that the model has captured semantic cues representing spatial location. By contrast, "in" shows lower attention towards the subject entity, aligning with the weaker type constraint of the "misc" label for the subject in this relation type.

Overall, the ROC model achieves hierarchical semantic modeling through its multi-layer encoder: shallow layers focus on context and global information, while deeper layers progressively concentrate on key entities and capture latent relational semantics. Additionally, the attention mechanism effectively filters redundant information, reducing interference from irrelevant words during training, thereby enhancing both the interpretability and robustness of the model.