Multimodal Document-level Triple Extraction via Dynamic Graph Enhancement and Relation-Aware Reflection

Xiang Li[1], **Runhai Jiao**[1*], **Changyu Zhou**[1], **Shoupeng Qiao**[1], **Ruojiao Qiao**[1], **Ruifan Li**[2]
[1]School of Control and Computer Engineering, North China Electric Power University, China
[2]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China
{120212227306,runhaijiao,zhouchangyu,120242227170,120242227168}@ncepu.edu.cn,
rfli@bupt.edu.cn

Abstract

Multimodal documents, which are among the most prevalent data formats, combine a large amount of textual and visual content. Extracting structured triples knowledge from these documents is a highly valuable task, aimed at helping users efficiently acquire key entities and their relationships. However, existing methods face limitations in simultaneously processing long textual content and multiple associated images for triple extraction. Therefore, we propose a Multimodal Document-level Triple Extraction (MDocTE) framework. Specifically, we introduce a dynamic document graph construction method that extends the model's scope to the entire document and the external world, while adaptively optimizing the graph structure. Next, we inject the global information and external knowledge learned by the graph neural network into the large language model, generating structured triples after deep interaction. Finally, we design a multimodal relation-aware mechanism and loss function to guide the model in reflecting on the shared information between text and visuals. We release a new triple extraction dataset for multimodal documents and conduct extensive experiments. The results demonstrate that the proposed framework outperforms the state-of-the-art baselines, thus filling the gap in multimodal document extraction. Our data is available at https://github.com/XiangLiphd/Triple-extraction-dataset.

1 Introduction

Triple extraction aims to identify entity pairs and their semantic relationships from unstructured text (Sun et al., 2024b; Ning et al., 2023; Naglik and Lango, 2024; Zhao et al., 2024a), providing a foundation for downstream tasks like knowledge graph construction, information retrieval, and knowledge enhancement (Wang et al., 2024a; Ovadia et al., 2024; Nguyen et al., 2024). Unlike the pipeline approach that separates named entity recognition and relation extraction, joint extraction integrates both tasks, mitigating error propagation and strengthening their intrinsic connections, making it a current research hotspot (Mullick et al., 2024; Shang et al., 2022; Huang et al., 2023).

However, most extraction datasets are sourced directly from vast and unstructured web text (He et al., 2023; Zheng et al., 2021). In many cases, the entities and relationships extracted directly from text are inaccurate (Cui et al., 2024; Zhang et al., 2025b). Thus, leveraging image information to correct extracted relationships becomes a promising strategy (Zheng et al., 2023; Hu et al., 2023; Chen et al., 2022). Yuan et al. (2023) proposed the first joint multimodal entity-relation extraction model, achieving fine-grained alignment between textual entities and visual objects. Building on this, yuan et al. (2024) further introduced a few-shot multimodal entity-relation extraction model, addressing the challenge of requiring large amounts of labeled data. These methods not only demonstrate that the incorporation of image information enhances the triple extraction model's ability to perceive textual semantic information (Wang et al., 2024c; Shen et al., 2024), but also provide us with strong support for document-level multimodal triple extraction using large language models (LLMs) (Hu et al., 2024; He et al., 2025). However, their datasets typically contain only short text segments paired with a single image, representing an idealized scenario.

It is worth mentioning that multimodal documents are among the most common data formats in everyday life, such as academic papers, news reports, social media posts, and more. However, traditional multimodal entity-relation extraction models face numerous challenges when processing multimodal documents. Firstly, many existing methods face challenges in capturing the global document structure and multimodal correlations, especially

[*]Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 3212–3223
November 4-9, 2025 ©2025 Association for Computational Linguistics

when the document is complex, leading to insufficient interaction of contextual multimodal data (Oral and Eryiğit, 2022; Kong et al., 2024). Although current language models can handle long text sequences and visual models can process multiple images, their joint processing capability significantly decreases when extended to documents containing multiple images and long texts. Achieving good extraction results requires not only capturing global information from the document but also thoroughly understanding the multimodal correlations (Xu et al., 2021; Zeng et al., 2020; Zhang et al., 2020). Our framework is designed based on this concept. Secondly, integrating a large amount of textual and visual information at the document level remains a complex task, as models must effectively align entities across modalities while ensuring semantic consistency (Hei et al., 2024). Lastly, while image modules can provide additional cues, efficiently leveraging multimodal knowledge to correct and refine extracted relations remains a significant challenge (Yuan et al., 2023). To the best of our knowledge, this task is highly challenging, and our proposed framework is the first attempt to address it.

To address these challenges, we propose a multimodal document-level triple extraction framework (MDocTE) via dynamic graph enhancement and multimodal relation-aware reflection. Specifically, it divides the document into multiple fine-grained modules and selectively associates them, primarily including global summaries, chunks, images, entities, visual objects, and external knowledge. It is worth noting that we have added feature interaction nodes between chunks and images, aiming to achieve fine-grained cross-modal information fusion. Additionally, in order to improve the extraction effectiveness and efficiency, we optimize the document graph based on the chunks to be extracted. Next, the knowledge learned by the graph neural network is deeply interacted with the LLMs to generate the raw entities and relationships. Finally, we design a multimodal relation-aware mechanism and loss function to prompt the large model to reflect on the shared information between text and visuals, generating accurate triple answers. The main contributions are summarized as follows:

(1) To the best of our knowledge, we propose the first triple extraction framework specifically designed for multimodal documents. It injects the global multimodal information learned and external knowledge learned by graph neural networks into the LLM, significantly improving extraction accuracy.

(2) We propose a dynamic document graph construction method. It leverages a divide-and-conquer strategy to hierarchically partition documents into multiple granularities, and enhances the richness of the document graph with external knowledge. Moreover, we introduce cross-modal feature interaction nodes for fine-grained multimodal fusion.

(3) We develop a multimodal relation-aware mechanism that prompts the LLM to reflect upon shared textual and visual information, mitigating incorrect inferences about complex relationships. Additionally, we propose a reflection loss that encourages the model to adjust its predictions by leveraging multimodal cues.

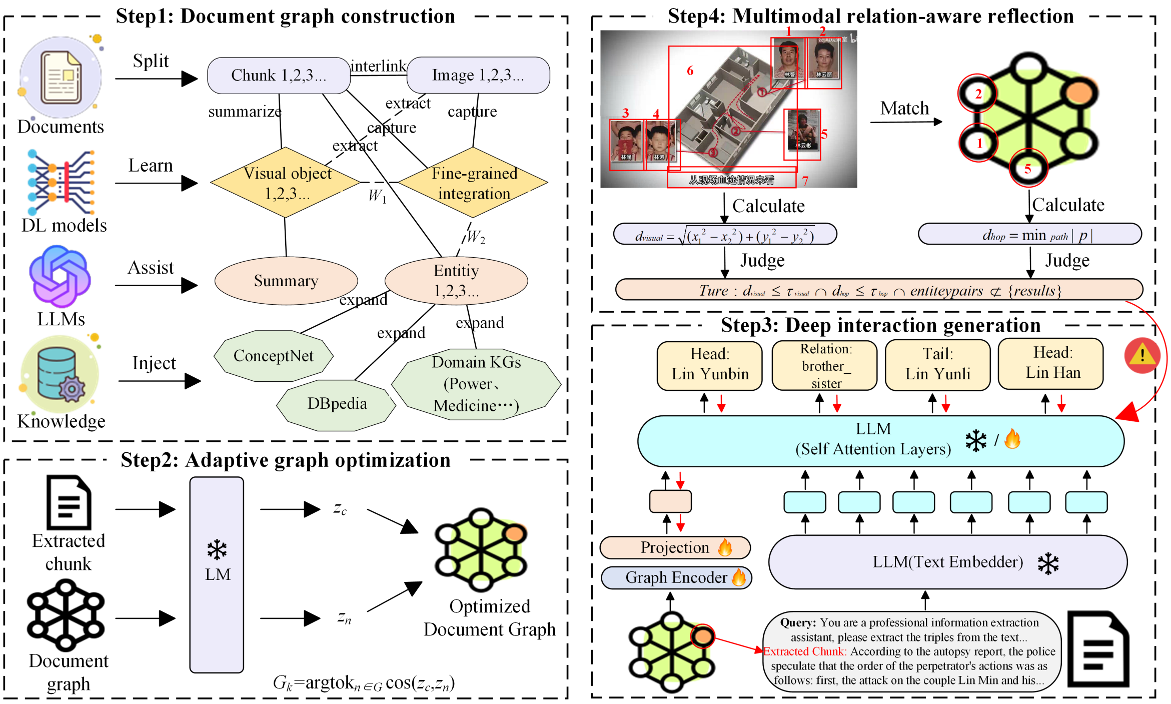
(4) We release a new triple extraction dataset for multimodal documents and conduct extensive experiments. Experimental results demonstrate that our framework outperforms existing models with significant improvements in precision (P), recall (R), and F1 scores.

2 Related Work

This section reviews the strengths and limitations of existing triple extraction methods from three perspectives: traditional, multimodal, and document-level triple extraction.

2.1 Triple extraction

Entity and relation extraction are fundamental tasks in information extraction. Early approaches typically handled named entity recognition (NER) and relation extraction (RE) separately through a pipeline (Li et al., 2022; Jehangir et al., 2023; Zhao et al., 2024b; Detroja et al., 2023). However, such methods suffer from error propagation and lack interaction between the tasks (Bekoulis et al., 2018; Qiao et al., 2022). Recent research has focused on joint models that tackle NER and RE simultaneously, improving the performance by capturing dependencies between the tasks (Gupta et al., 2016; Xu et al., 2022). However, such models are typically limited to short texts and face challenges in handling complex scenarios with rich multimodal information or large-scale documents (Zhang et al., 2024).



Query: You are a professional information extraction assistant, please extract the triples from the text... Extracted Chunk: According to the autopsy report, the police speculate that the order of the perpetrator's actions was as follows: first, the attack on the couple Lin Min and his...

Figure 1: The entire extracting process of proposed framework

2.2 Multimodal triple extraction

In recent years, with the widespread rise of multimodal generative models, leading to a growing interest in multimodal triple extraction (MTE)(Yang et al., 2023). Specifically, Yuan et al. (2023) designed an edge-enhanced graph alignment network that enhances MTE by aligning the nodes and edges across cross-modal graphs. Building on this, Wang et al. (2024b) proposed a temporal enhancement and similarity-gated attention network for MTE, which alleviated the issue of irrelevant information interference in images. Additionally, Yuan et al. (2025) further proposed a fine-grained multimodal network to tackle limitations in cross-modal fine-grained alignment, irrelevant information filtering, and multi-word entity extraction. These methods have advanced the enhancement of textual semantics with visual information, but challenges persist in handling large-scale documents, particularly in capturing long-range dependencies and integrating multimodal information across multiple images and texts(Cao and Wu, 2023).

2.3 Document-level triple extraction

Document-level triple extraction (DTE) is essential for real-world applications where information spans entire documents (Li et al., 2023; Meng et al., 2023). Recent research mainly relies on graph-based and large language models-based approaches. Specifically, Zhang et al. (2023) reformulated DTE as a table-filling task and proposed a table-to-graph generation model to enable fine-grained information sharing. In addition, Sun et al. (2024a) and Zhang et al. (2025a) adapted large language models for DTE, addressing the issue of context forgetting. These methods have been successful in DTE, but they lack sufficient integration of external knowledge and struggle to handle multimodal data. To address these issues, Liu et al. (2024) constructed a multimodal, multilingual and multitask dataset for grounded DTE, and designed a hierarchical multimodal information extraction model. It is worth noting that this research opens up new directions for us, but the constructed dataset contains relatively short texts, which makes it difficult to match the document length found in real-world scenarios.

3 Approach

Building on these foundations, we propose the first triple extraction framework designed specifically for multimodal documents (MDocTE). As shown in Figure 1, the proposed framework consists of four main steps: document graph construction, adaptive graph optimization, deep interaction generation, and multimodal relation-aware reflection. The implementation details of each step are presented in the following sections.

3.1 Document graph construction

To better understand the content of multimodal documents, we hierarchically divide them into multiple fine-grained modules, primarily consisting of chunks, images, entities, and visual objects. These modules exhibit hierarchical relationships, such as interlink relationships between chunks and images, and extraction relationships between chunks and entities, as well as between images and visual objects. It is worth noting that entities and visual objects are extracted using pre-trained models. Additionally, to refine the document's key content and eliminate irrelevant details, we employ a large language model to generate summary nodes, which are then linked to the chunks.

To better aggregate text and image information, we propose a fine-grained feature selection module. It is worth noting that it further deepens the relationship between associated entities and visual object pairs based on the knowledge-enhanced network proposed by Huang et al. (2025). Specifically, given a chunk-image pair, we first use a pre-trained model to extract entities and visual objects, and generate fine-grained features. The entities are extracted using Deepseek and encoded with the BERT-base model. Visual objects with the top-k confidence scores are identified using a pre-trained Mask-RCNN (He et al., 2017) model, and the corresponding features are extracted. The encoding process for entities and visual objects is illustrated in equations (1) and (2).

$$E = \mathrm{BERT}(e) = [e_1, e_2, \ldots, e_n] \tag{1}$$

$$V = \mathrm{Mask-RCNN}(i) = [v_1, v_2, \ldots, v_k] \tag{2}$$

Next, we use equations (3-4) to identify the visual object v that is highly relevant to each entity and assign a higher weight to v.

$$\alpha_{ij}^T = \mathrm{softmax}(W_3 \tanh(W_1 v_j + W_2 e_i)) \tag{3}$$

$$m_{e_i} = \sum_{j \in k} \alpha_{ij}^T v_j \tag{4}$$

Here, W represents the parameter matrix. Subsequently, we generate the enhanced entities features based on equations (5-6). Similarly, we generate the enhanced visual object representation e^*. Finally, we use cross-attention to generate fine-grained interaction node feature representations.

$$f_i^T = \mathrm{sigmoid}(W_4[e_i; m_{e_i}]) \tag{5}$$

$$e_i^* = e_i + f_i^T m_{e_i} \tag{6}$$

Considering that some visually similar objects in the image may mislead Mask-RCNN into generating incorrect feature representations, we employ contrastive learning to differentiate them. Specifically, we calculate the similarity between entities and visual objects using CLIP (Radford et al., 2021) or a domain-specific fine-tuned version of CLIP. We then select difficult sample pairs (with the highest similarity treated as positive examples, and those with similarity greater than 0.5 as negative examples) for distinction and learning. This approach aims to help Mask-RCNN generate more accurate vector representations. It is worth noting that in contrastive learning, we use the commonly used NT-Xent loss function.

It is worth mentioning that, to broaden the scope of the extraction model, we incorporate external knowledge bases such as ConceptNet, DBpedia, or domain knowledge graphs into the document graph. Specifically, we search for all the entities appearing in the knowledge base and connect their adjacent knowledge to the document graph, with the expand relationship.

3.2 Adaptive graph optimization

To adopt the same encoding strategy as the generative model, we use the BERT to encode the text in the document graph. Additionally, for the images in the document graph, we use the CLIP encoder mentioned in the previous section to generate vector representations. Next, to identify the nodes most relevant to the current extraction task, we calculate the distances between nodes using cosine similarity and filter out irrelevant nodes using k-nearest neighbor retrieval. The detailed process is shown in equation (7). It is worth noting that, in order to retain as much useful information as possible while filtering out complex text, we first filter the one-hop neighboring nodes of the target extraction chunk node, then filter the two-hop neighboring nodes, and finally use the remaining optimized subgraph vectors as input to the graph neural network in the next section. This step offers two key benefits: first, it helps filter out nodes and edges that are unrelated to the current extraction task. If the entire document graph were input into the graph neural network as augmentation information for the LLM without any filtering, it could distract the LLM from the relevant extraction information or even cause interference. Second, pruning the graph significantly enhances extraction efficiency, which is crucial.

$$G_k = \text{argtok}_{n \in G} \cos(z_c, z_n) \tag{7}$$

where z_c and z_n are the embeddings of target extraction node and other nodes.

3.3 Deep interaction generation

To acquire the global multimodal information and external knowledge of the chunk to be extracted, we use a graph attention neural network (GAT) to learn the optimized document graph and update the node representation:

$$h_c = \text{GNN}(G_c) \in \mathbb{R}^{d_g} \tag{8}$$

Here, d_g represents the dimension of the graph encoder. Next, to align the graph tokens with the vector space of the LLM, we utilize a multi-layer perceptron (MLP):

$$h_c^* = \text{MLP}(h_c) \in \mathbb{R}^{d_l} \tag{9}$$

where d_l is the dimension of the LLM's hidden embedding. It is worth noting that the parameters of the graph neural network and the MLP are not frozen and are updated throughout the task. Additionally, the fine-grained feature interaction parameters are updated within the graph neural network.

For the input query, we generate their representation vectors ht using a text embedder, which is the first layer of the LLM. Finally, we input both the node representation of chunk and query vectors into the LLM to preliminarily generate the triple answers. The generation process is as follows:

$$p_{\theta,\phi_1,\phi_2}(Y \mid x_t) = \prod_{i=1}^{r} p_{\theta,\phi_1,\phi_2}\left(y_i \mid y_{<i}, [h_c^*; h_t]\right) \tag{10}$$

where θ represents the parameters of the LLM, while ϕ_1 and ϕ_2 represent the parameters of the graph neural network and the MLP, respectively. During triple extraction, we use the cross-entropy loss function for model training.

3.4 Multimodal relation-aware reflection

To improve the understanding of complex relationships in LLM and generate complete triples, we designed a multimodal relation-aware mechanism that prompts the model to deeply reflect on pseudo labels. The following details the generation process of pseudo labels. It is well known that there is often a relationship between adjacent or nearby visual objects in images. Similarly, the same applies to text. Therefore, we use equations (11) and (12) to calculate the distance between visual objects in the image and the distance between entities.

$$d_{\text{visual}} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{11}$$

$$d_{\text{hop}} = \min_{\text{path}} |p| \tag{12}$$

Here, d_{visual} represents the distance between the center points of visual object bounding boxes detected by Mask-RCNN, and d_{hop} represents the shortest path distance between two entities in the document graph. It is worth noting that for the matching process between visual objects and entities, we use CLIP to calculate the similarity and select the most similar ones that exceed a certain threshold. If a pair of visual objects can be matched to a corresponding pair of entities, and both d_{visual} and d_{hop} are below a certain threshold, we consider that there is a relationship between these two entities (pseudo label).

Next, we check whether the generated results of the LLM contain pseudo label entity pairs. If there are missing relationships, we prompt the model to reflect on them. Additionally, we design a reflection loss to guide the model in deeper reasoning as shown in equation (13). The reason for this design is that LLM often confidently output ‘no relation’ when dealing with implicit or uncertain relationships, and this loss can encourage them to better predict the relationships between entity pairs.

$$L_{\text{reflect}} = -\frac{1}{|\varphi|} \sum_{(e_a, e_b) \in \varphi} \log\left(1 - p_{\text{no-rel}}(e_a, e_b)\right) \tag{13}$$

where φ is the set of pseudo labels, and $p_{\text{no-rel}}$ represents the probability of generating no relation.

4 Experiment

4.1 Dataset and evaluation metrics

To evaluate the effectiveness of our framework, we modified the data from Liu et al. (2024). The preprocessed dataset involved splitting a long video into multiple segments and extracting both images and text data. Considering that most documents are much longer than this dataset, we performed effective integration and added more distant labels to better simulate real-world scenarios. To the best of our knowledge, this is the only dataset for triple

Categories	Methods	P	R	F1
LLM	Llama-3-8b-Chinese	12.88%	13.57%	13.22%
	GPT-4o	14.14%	20.58%	16.76%
	Deepseek	16.2%	19.22%	17.58%
	Llama-3-8b+LoRA	24.58%	26.63%	25.58%
DTEM	BERT+LSTM+CRF (Zheng et al., 2017)	15.88%	16.34%	16.11%
	Onerel (Shang et al., 2022)	19.57%	22.79%	21.06%
	ZeroDocRTE (Sun et al., 2024a)	18.62%	22.56%	20.4%
	RAKG (Zhang et al., 2025a)	16.53%	17.92%	17.2%
MTEM	HVPNeT (Chen et al., 2022)	31.98%	36.44%	34.07%
	PROMU (Hu et al., 2023)	30.62%	33.11%	31.81%
	TMR (Zheng et al., 2023)	33.77%	37.8%	35.68%
	EEGA (Yuan et al., 2023)	19.55%	21.35%	20.41%
	TESGA (Wang et al., 2024b)	18.29%	20.07%	19.13%
DMTEM	M3D (Liu et al., 2024)	23.05%	24.9%	23.94%
	MDocTE (Llama-3-8b)	27.13%	31.62%	29.2%
	MDocTE (Qwen-8b)	28.0%	32.95%	30.27%
	MDocTE+LoRA (Llama-3-8b)	30.87%	38.24%	34.16%
	MDocTE+LoRA (Qwen-8b)	32.47%	40.12%	35.89%

Table 1: Performances of Triple Extraction Methods

extraction in multimodal documents. Specifically, the dataset contains 982 documents, 11891 images, and 54993 triples. During the training process, we split the dataset into training, validation, and test sets in an 8:1:1 ratio. Furthermore, we employ the most widely used evaluation metrics in this task: Precision (P), Recall (R), and F1 score.

4.2 Implementation details

All experiments are performed on a Linux-based server with 2 NVIDIA A40 GPUs. The graph encoder, i.e. GAT, has 4 layers with 4 heads per layer and a hidden dimension size of 1024. We use Llama-3-8b as the backbone for the large language model, and fine-tune it with low-rank adaption (LoRA). Moreover, the main parameters are set for the model include Epoch as 30, dropout as 0.2, Optimizer as Adam, batch_size as 2, and Learning_rate as 1×10^{-5}. In adaptive graph optimization, k can be set to 3, 5, or 10.

4.3 Baselines

To demonstrate the effectiveness of proposed framework, we compare its performance to widely used models on triple extraction task. They are classified into the following four categories: commonly used large language models, document-level triple extraction models (DTEM), multimodal triple extraction models (MTEM), and document-level multimodal triple extraction models (DMTEM). Among them, the LLMs include Llama3-8b-Chinese, GPT-4o, and DeepSeek, which are the most widely used currently. BERT+LSTM+CRF (Zheng et al., 2017) and Onerel (Shang et al., 2022) are the most classical DETM. Additionally, ZeroDocRTE (Sun et al., 2024a) and RAKG (Zhang et al., 2025a) implement document-level triple extraction based on large language models, demonstrating their advantages. HVPNeT (Chen et al., 2022) uses visual representations as pluggable prefixes to guide textual representations, laying a solid foundation for MTEM. Moreover, PROMU (Hu et al., 2023), TMR (Zheng et al., 2023), and EEGA (Yuan et al., 2023) have been extensively applied in MTE, effectively achieving image-text alignment through the innovative design of distinct architectural frameworks. Building on this, TESGA (Wang et al., 2024b) better integrates image and text information, demonstrating superior performance. M3D (Liu et al., 2024) is one of the few existing models for DMTEM, laying a solid foundation for subsequent research.

4.4 Main results

Table 1 presents the experimental results of existing triple extraction methods across different categories. It is important to note that HVPNeT, PROMU, and TMR require entity information as input prior to extraction, and thus are not per-forming genuine end-to-end triple extraction. Overall, MDocTE achieves state-of-the-art performance across multiple evaluation metrics, demonstrating that the proposed dynamic document graph and relation-aware reflection mechanism effectively enhance the triple extraction capabilities of LLMs. Specifically, even without LoRA fine-tuning, MDocTE surpasses current LLMs, indicating that existing LLMs lack sufficient global document comprehension and fine-grained insight. We observe that Deepseek considers complex inter-entity relationships during deeper reasoning but often omits them in the final output due to uncertainty about their correctness or user requirement. However, after incorporating our reflection prompting strategy, it develops new insights into these entities and generates the correct triples, thereby validating the effectiveness of the proposed reflection mechanism. It is worth noting that MDocTE with LoRA fine-tuning achieves additional performance gains, further confirming the framework's effectiveness and scalability. Moreover, experimental results show that existing DTEM and MTEM methods struggle to effectively utilize long text and multiple image information. On the contrary, MDocTE outperforms other triple extraction methods, benefiting from the injected global information and external knowledge.

4.5 Ablation study

Methods	F1	Decline
w/o Graph encoder	19.86%	9.34%
w/o Fine-grained feature selection	26.88%	2.32%
w/o External knowledge	27.15%	2.05%
w/o Reflection	24.67%	4.53%

Table 2: Ablation experiments of proposed framework

Table 2 presents the experimental results of removing each improvement component. It is evident that the graph encoder has the most significant impact on the extraction performance, demonstrating that incorporating hierarchical information from document graphs can enhance the model's ability to understand extraction tasks. Secondly, removing the reflection mechanism led to a 4.53% decline in performance, highlighting its role in helping the model resolve complex relationships in multimodal data, such as implied and long-distance relationships. It should be noted that Section 4.9 provides more vivid demonstrations of the multimodal reflection mechanism's functionality. Additionally, the extraction performance of LLM was adversely affected when removing either the fine-grained feature selection module or external knowledge sources during document graph construction. This further confirms that building a comprehensive document graph enables LLM to acquire enhanced knowledge representation, thereby improving extraction effect.

4.6 Mitigation of hallucination

Methods	Number of hallucinations
Llama-3-8b-Chinese	53
GPT-4o	0
Deepseek	2
Ours	5

Table 3: Hallucination analysis experiment

Table 3 shows the number of hallucinated triples that appeared during testing with different large language models. It should be noted that hallucinations in triple extraction refer to entities not present in the document or relationships that are not within the given types. Clearly, compared to Llama3, the improved framework essentially eliminates hallucinations and reaches a level comparable to GPT and DeepSeek, which demonstrates that the proposed framework can understand document knowledge and perform self-reflection.

4.7 Efficiency evaluation

Methods	F1	Time/Epoch
Before graph optimization	27.49%	135
After graph optimization	29.2%	106

Table 4: Efficiency evaluation experiments

Table 4 shows the experimental results before and after document graph optimization. It is obvious that not only the F1 value is improved by 1.71% after the optimization, the average training time per epoch is also decreased by 29 minutes. This proves that our optimization using extracted chunks is effective, not only filtering information unrelated to the content of this extraction, but also greatly improving the efficiency.

Methods	P	R	F1
Llama-3-8b-Chinese	24.34%	25.69%	24.99%
Llama-3-8b+LoRA	55.31%	50.26%	52.67%
MDocTE+LoRA	59.28%	55.41%	57.28%

Table 5: Performances on Power Domain Dataset

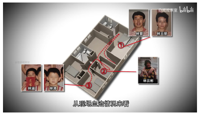
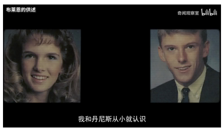
Case 1		Case 2	
Images: 林云丽 Lily 谢连斌 Robert 林姝 Kathy 女婿谢连斌和女儿林姝 **(a)**  **(b)**	**Text:** ...the five victims are 45-year-old homeowner Lin Min, his 43-year-old wife Lin Yunli, his 39-year-old sister Lin Yunbin... **Golden:** (Lin Min, PER-PER_brother_sister, Lin Yunbin), (Lin Yunli, PER-PER_brother_sister, Lin Yunbin)	**Images:**  **(a)** **(b)**	**Text:** ...Dennis and I have known each other since we were kids, and then Dennis got married and we still dated closely... **Golden:** (Dennis, PER-PER_peer, Brian), (Dennis, PER-PER_couple, Brian)
Llama3-8b: (Lin Min, PER-PER_peer, Lin Yunbin), (Lin Yunli, no relation, Lin Yunbin) ✗		**Llama3-8b:** (Dennis, no relation, Brian) ✗	
GPT-4: (Lin Min, PER-PER_brother_sister, Lin Yunbin), (Lin Yunli, PER-PER_peer, Lin Yunbin) ✗		**GPT-4:** (Dennis, PER-PER_peer, Brian) ✗	
Deepseek: (Lin Min, PER-PER_brother_sister, Lin Yunbin), (Lin Yunli, no relation, Lin Yunbin) ✗		**Deepseek:** (Dennis, PER-PER_peer, Brian) ✗	
Ours: (Lin Min, PER-PER_brother_sister, Lin Yunbin), (Lin Yunli, PER-PER_brother_sister, Lin Yunbin) ✓		**Ours:** (Dennis, PER-PER_peer, Brian), (Dennis, PER-PER_couple, Brian) ✓	

Table 6: Case study experiment

4.8 Domain applicability

To further validate the generalizability of the framework, we conducted domain-specific experiments on power dataset, which comprises 1043 documents (e.g., maintenance logs and fault reports) and 1901 images. As shown in Table 5, compared to the baseline model, our framework achieves improvements of 3.97%, 5.15%, and 4.61% in P, R, and F1, demonstrating cross-domain adaptation capability. It should be noted that this enhancement is partly attributed to our introduction of the domain knowledge base constructed in previous studies.

4.9 Case study

Table 6 vividly illustrates the experimental results of two cases. Specifically, in Case 1, Figure (a) shows a family photo of Lin Min, and Figure (b) shows the house where the victims were found. It is worth noting that in the top right corner of Figure (b) are Lin Min and his wife, and directly below them is Lin Min's sister, Lin Yunbin. The most challenging relationship to identify in this case is the relationship between Lin Min's wife, Lin Yunli, and Lin Min's sister, Lin Yunbin. Both Llama3 and DeepSeek clearly misinterpret them as unrelated. Additionally, due to the frequent mention of ages in the text, GPT-4 rustily assumes they are peers. However, in reality, their relationship is indirect and implied within the familial context. The output of the proposed framework is correct, demonstrating that the multimodal relation-aware mechanism can identify some pseudo labels and prompt the large language model to self-correct. Case 2 presents a more difficult relationship determination. Figure (a) depicts a childhood photo of Dennis and Brian, whereas Figure (b) is a wedding photo of Dennis and Mike. Based on the text and images, it is clear that Dennis and Brian were childhood friends, which most models also assume. However, a deeper understanding of the document reveals that Dennis and Brian killed Mike and maintained an affair, introducing an additional couple relationship. Only our model successfully captures this relationship, demonstrating its ability to fully comprehend the contextual content of the document. Furthermore, applying reflection prompts to DeepSeek and other large models can enable them to revise their original answers as well, further demonstrating the effectiveness of the multimodal relation-aware mechanism.

5 Conclusion

In this paper, we have integrated a multimodal document extraction dataset and designed a triple extraction framework. It not only expands the scope to global document context and external world but also leverages multimodal shared information to prompt the model for relation-aware reflection. Experimental results demonstrate that the proposed framework outperforms existing methods and effectively mitigates hallucinations. In the future, we will expand more datasets to validate the effectiveness of the proposed framework.

Limitations

Currently, although the experimental results of the proposed framework have surpassed existing methods, there is still significant room for improvement. For example, the current framework only considers the alignment between a single text and multiple images, without considering the alignment between multiple images. In the future, more complex triplet extraction frameworks will be explored to address the practical extraction challenges in multimodal documents. In addition, the current dataset is relatively monolingual.

Acknowledgements

We are grateful to the anonymous reviewers and the area chair for their insightful comments and suggestions. This research was supported by the National Natural Science Foundation of China (Grant No. 62373150).

References

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. Expert Systems with Applications, 114:34–45.

Panfeng Cao and Jian Wu. 2023. Graphrevisedie: Multimodal information extraction with graph-revised network. Pattern Recognition, 140:109542.

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. 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, Seattle, United States. Association for Computational Linguistics.

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.

Kartik Detroja, C.K. Bhensdadia, and Brijesh S. Bhatt. 2023. A survey on relation extraction. Intelligent Systems with Applications, 19:200244.

Pankaj Gupta, Hinrich Schütze, and Bernt Andrassy. 2016. Table filling multi-task recurrent neural network for joint entity and relation extraction. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2537–2547, Osaka, Japan. The COLING 2016 Organizing Committee.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 2980–2988.

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, MM '23, page 4564–4573, New York, NY, USA. Association for Computing Machinery.

Xinyu He, Shixin Li, Yuning Zhang, Binhe Li, Sifan Xu, and Yuqing Zhou. 2025. The more quality information the better: Hierarchical generation of multi-evidence alignment and fusion model for multimodal entity and relation extraction. Information Processing Management, 62(1):103875.

Lei Hei, Ning An, Tingjing Liao, Qi Ma, Jiaqi Wang, and Feiliang Ren. 2024. Multimodal relational triple extraction with query-based entity object transformer.

Huiyun Hu, Junda Kong, Fei Wang, Hongzhi Sun, Yang Ge, and Bo Xiao. 2024. Gmner-lf: Generative multimodal named entity recognition based on llm with information fusion. In 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pages 1–6.

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, MM '23, page 5185–5194, New York, NY, USA. Association for Computing Machinery.

Qing Huang, Yanbang Sun, Zhenchang Xing, Min Yu, Xiwei Xu, and Qinghua Lu. 2023. Api entity and relation joint extraction from text via dynamic prompt-tuned language model. ACM Trans. Softw. Eng. Methodol., 33(1).

Shubin Huang, Yi Cai, Li Yuan, and Jiexin Wang. 2025. A knowledge-enhanced network for joint multimodal entity-relation extraction. Information Processing Management, 62(3):104033.

Basra Jehangir, Saravanan Radhakrishnan, and Rahul Agarwal. 2023. A survey on named entity recognition — datasets, tools, and methodologies. Natural Language Processing Journal, 3:100017.

Lingxing Kong, Jiuliang Wang, Zheng Ma, Qifeng Zhou, Jianbing Zhang, Liang He, and Jiajun Chen. 2024. A hierarchical network for multimodal document-level relation extraction. Proceedings of the AAAI Conference on Artificial Intelligence, 38(16):18408–18416.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1):50–70.

Junpeng Li, Zixia Jia, and Zilong Zheng. 2023. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 5495–5505, Singapore. Association for Computational Linguistics.

Jiang Liu, Bobo Li, Xinran Yang, Na Yang, Hao Fei, Mingyao Zhang, Fei Li, and Donghong Ji. 2024. M^3d: A multimodal, multilingual and multitask dataset for grounded document-level information extraction.

Shiao Meng, Xuming Hu, Aiwei Liu, Shuang Li, Fukun Ma, Yawen Yang, and Lijie Wen. 2023. RAPL: A relation-aware prototype learning approach for few-shot document-level relation extraction. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 5208–5226, Singapore. Association for Computational Linguistics.

Ankan Mullick, Sombit Bose, Abhilash Nandy, Gajula Sai Chaitanya, and Pawan Goyal. 2024. A pointer network-based approach for joint extraction and detection of multi-label multi-class intents. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 15664–15680, Miami, Florida, USA. Association for Computational Linguistics.

Iwo Naglik and Mateusz Lango. 2024. ASTE-transformer: Modelling dependencies in aspect-sentiment triplet extraction. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 2324–2339, Miami, Florida, USA. Association for Computational Linguistics.

Thong Nguyen, Shubham Chatterjee, Sean MacAvaney, Iain Mackie, Jeff Dalton, and Andrew Yates. 2024. DyVo: Dynamic vocabularies for learned sparse retrieval with entities. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 767–783, Miami, Florida, USA. Association for Computational Linguistics.

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.

Berke Oral and Gülşen Eryiğit. 2022. Fusion of visual representations for multimodal information extraction from unstructured transactional documents. International Journal on Document Analysis and Recognition, 25(3):187–205. Publisher Copyright: © 2022, The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-tuning or retrieval? comparing knowledge injection in LLMs. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 237–250, Miami, Florida, USA. Association for Computational Linguistics.

Bo Qiao, Zhuoyang Zou, Yu Huang, Kui Fang, Xinghui Zhu, and Yiming Chen. 2022. A joint model for entity and relation extraction based on bert. Neural Comput. Appl., 34(5):3471–3481.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Yu-Ming Shang, Heyan Huang, and Xianling Mao. 2022. Onerel: Joint entity and relation extraction with one module in one step. Proceedings of the AAAI Conference on Artificial Intelligence, 36(10):11285–11293.

Qianru Shen, Hailun Lin, Huan Liu, Zheng Lin, and Weiping Wang. 2024. Exploiting visual relation and multi-grained knowledge for multimodal relation extraction. In 2024 International Joint Conference on Neural Networks (IJCNN), pages 1–8.

Qi Sun, Kun Huang, Xiaocui Yang, Rong Tong, Kun Zhang, and Soujanya Poria. 2024a. Consistency guided knowledge retrieval and denoising in llms for zero-shot document-level relation triplet extraction.

Qiao Sun, Liujia Yang, Minghao Ma, Nanyang Ye, and Qinying Gu. 2024b. MiniConGTS: A near ultimate minimalist contrastive grid tagging scheme for aspect sentiment triplet extraction. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 2817–2834, Miami, Florida, USA. Association for Computational Linguistics.

Fali Wang, Runxue Bao, Suhang Wang, Wenchao Yu, Yanchi Liu, Wei Cheng, and Haifeng Chen. 2024a. InfuserKI: Enhancing large language models with knowledge graphs via infuser-guided knowledge integration. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 3675–3688, Miami, Florida, USA. Association for Computational Linguistics.

Guoxiang Wang, Jin Liu, Jialong Xie, Zhenwei Zhu, and Fengyu Zhou. 2024b. Joint multimodal entity-relation extraction based on temporal enhancement and similarity-gated attention. Knowledge-Based Systems, 304:112504.

Ziqi Wang, Chen Zhu, Zhi Zheng, Xinhang Li, Tong Xu, Yongyi He, Qi Liu, Ying Yu, and Enhong Chen. 2024c. Granular entity mapper: Advancing fine-grained multimodal named entity recognition and grounding. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 3211–3226, Miami, Florida, USA. Association for Computational Linguistics.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yabing Shi, Yong Zhu, Jie Gao, and Zhendong Mao. 2022. EmRel: Joint representation of entities and embedded relations for multi-triple extraction. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 659–665, Seattle, United States. Association for Computational Linguistics.

Runxin Xu, Tianyu Liu, Lei Li, and Baobao Chang. 2021. Document-level event extraction via heterogeneous graph-based interaction model with a tracker. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3533–3546, Online. Association for Computational Linguistics.

Li Yang, Jieming Wang, Jin-Cheon Na, and Jianfei Yu. 2023. Generating paraphrase sentences for multimodal entity-category-sentiment triple extraction. Knowledge-Based Systems, 278:110823.

li yuan, Yi Cai, and Junsheng Huang. 2024. Few-shot joint multimodal entity-relation extraction via knowledge-enhanced cross-modal prompt model. In Proceedings of the 32nd ACM International Conference on Multimedia, MM '24, page 8701–8710, New York, NY, USA. Association for Computing Machinery.

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 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.

Li Yuan, Yi Cai, Jingyu Xu, Qing Li, and Tao Wang. 2025. A fine-grained network for joint multimodal entity-relation extraction. IEEE Transactions on Knowledge and Data Engineering, 37(1):1–14.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1630–1640, Online. Association for Computational Linguistics.

Hairong Zhang, Jiaheng Si, Guohang Yan, Boyuan Qi, Pinlong Cai, Song Mao, Ding Wang, and Botian Shi. 2025a. Rakg:document-level retrieval augmented knowledge graph construction.

Meishan Zhang, Hao Fei, Bin Wang, Shengqiong Wu, Yixin Cao, Fei Li, and Min Zhang. 2024. Recognizing everything from all modalities at once: Grounded multimodal universal information extraction. In Findings of the Association for Computational Linguistics: ACL 2024, pages 14498–14511, Bangkok, Thailand. Association for Computational Linguistics.

Ruoyu Zhang, Yanzeng Li, and Lei Zou. 2023. A novel table-to-graph generation approach for document-level joint entity and relation extraction. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10853–10865, Toronto, Canada. Association for Computational Linguistics.

Ying Zhang, Gaoxiang Li, Hu Gao, and Depeng Dang. 2025b. Multi-scale interaction network for multimodal entity and relation extraction. Information Sciences, 699:121787.

Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Tingwen Liu, Hengzhu Tang, Wang Yubin, and Li Guo. 2020. Document-level relation extraction with dual-tier heterogeneous graph. In Proceedings of the 28th International Conference on Computational Linguistics, pages 1630–1641, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jianli Zhao, Changhao Xu, and Bin. Jiang. 2024a. IPED: An implicit perspective for relational triple extraction based on diffusion model. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2080–2092, Mexico City, Mexico. Association for Computational Linguistics.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024b. A comprehensive survey on relation extraction: Recent advances and new frontiers. ACM Comput. Surv., 56(11).

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, Toronto, Canada. Association for Computational Linguistics.

Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021. 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.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1227–1236, Vancouver, Canada. Association for Computational Linguistics.