JMERE-R1: Reasoning Enhanced LVLMs for Joint Multimodal Entity-Relation Extraction

Anonymous ACL submission

Abstract

Joint Multimodal Entity and Relation Extraction (JMERE) aims to extract structured entity-relation quintuplets from textual sequences with social media images. Large Vision-Language Models (LVLMs) demonstrate impressive performance across various multimodal downstream tasks. However, due to the complexity of quintuple extraction logic and multimodal information fusion, higher demands are placed on the model's ability to capture associations between modalities and perform reasoning. Current LVLMs still perform poorly on the JMERE task. To address these challenges, we propose JMERE-R1, a novel reasoning-enhanced paradigm for LVLMs. Our method integrates Supervised Fine-Tuning (SFT) with Reinforcement Learning (RL) to guide LVLMs toward autonomous reasoning in multimodal contexts. Furthermore, we employ automatically generated Multimodal Paradigm Chain-of-Thought (MP-CoT) data to encourage the model to focus more on Image and text interaction information. Experimental results show that with parameter-efficient fine-tuning and reinforcement learning, the LVLM is able to develop autonomous multimodal reasoning capabilities. Combined with our Policy-guided approach for multimodal information capture and association, JMERE-R1 enables the LVLM to achieve significantly stronger performance.

1 Introduction

Joint Multimodal Entity and Relation Extraction (JMERE) aims to extract entity–relation quintuplets from short text sequences accompanied by auxiliary images on social media platforms (Yuan et al., 2023). This task unifies Multimodal Named Entity Recognition (MNER) and Multimodal Relation Extraction (MRE), enabling the model to leverage the inherent correlations between them (Lu et al., 2018; Zheng et al., 2021b,a). Previous approaches to multimodal information extraction predominantly rely on discriminative models combined with image-text alignment modules. Recently, Large Vision-Language Models (LVLMs) (Yang et al., 2024) composed of visual encoders, cross-modal projectors, and large language models (LLMs) demonstrate strong generalization and multimodal reasoning capabilities after pretraining on massive multimodal corpora. These advancements raise the question: Can LVLMs be more effectively leveraged to improve JMERE performance by harnessing their powerful pretraining knowledge and reasoning abilities?

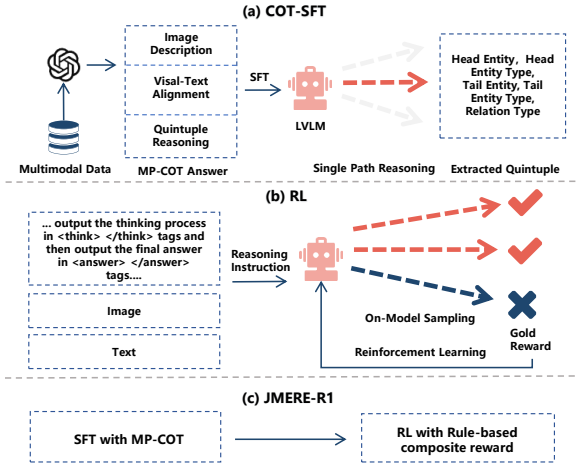

Figure 1: Three reasoning-enhanced paradigms. Gray arrows indicate potential reasoning paths, red indicates the correct path, and blue indicates incorrect paths.

However, our experiments show that applying conventional Supervised Fine-Tuning (SFT) to LVLMs for JMERE tasks has notable limitations (Trung et al., 2024; Fu et al., 2024). First, standard SFT only fits token-level probabilities to correct outputs, without directly optimizing for multimodal understanding or structured output prediction, which limits performance. Second, the reasoning behind multimodal inference is opaque, making the extracted quintuplets hard to interpret and hindering effective error analysis and targeted improvements.

Chain-of-Thought (CoT) prompting has proven effective in eliciting reasoning from LLMs in tasks such as arithmetic and commonsense reasoning. We posit that similar latent reasoning processes exist in JMERE, and that enabling LVLMs with explicit multimodal reasoning can significantly improve performance. Yet, to the best of our knowledge, this direction remains underexplored. As shown in Figure 1, we design and investigate three CoT-enhanced learning paradigms: Chain-of-Thought Supervised Fine-Tuning (CoT-SFT), Reinforcement Learning (RL) and Reasoning Enhanced JMERE (JMERE-R1). To avoid the high cost associated with acquiring additional human-annotated CoT data, we adopt a strategy that generates CoT supervision without requiring extra manual annotations. Specifically, we design a specialized format called Multimodal-Paradigm CoT (MP-CoT), which emphasizes the reasoning process involving visual understanding and image-text alignment within LVLMs. Each CoT follows a fixed three-part template: (1) Image Description, (2) Visal-Text Alignment, and (3) Quintuple Reasoning.

To construct MP-CoT data, we provide input samples and gold label to multimodal LLM, which generates reasoning paths based on a predefined format, with a dual-filtering mechanism to enhance CoT quality. The generated CoT samples are then used for parameter-efficient fine-tuning (PEFT) (Hu et al.; Dettmers et al., 2023) of the target LVLM. However, the CoT-SFT paradigm suffers from fixed reasoning paths, relying on the provided CoT annotations, which typically involve a single reasoning path. Notably, in JMERE, there are often multiple valid reasoning paths and information sources for extracting the same quintuple. This motivates us to adopt reinforcement learning techniques that enhance textual reasoning in LLMs (Liu et al., 2024). By comparing multiple generated reasoning paths and rewarding the best ones, we encourage the model to develop more diverse and accurate reasoning behaviors.

In our further exploration, we test whether reinforcement learning alone can stimulate the reasoning abilities of LVLMs. At each training step, we use the GPRO algorithm (Shao et al., 2024) to update the model. By comparing multiple reasoning outputs sampled for the same query with the gold label, we provide directional rewards to guide learning. This paradigm focuses on encouraging autonomous reasoning within LVLMs. However, since untrained LVLMs lack prior awareness of the JMERE task format, their initial accuracy is very low, making convergence difficult. As a result, reinforcement learning alone struggles to provide strong JMERE performance.

To address the limitations of both conventional SFT and the above reasoning-inspired RL paradigm, we propose a novel training framework called JMERE-R1. This approach combines CoT-based fine-tuning with reinforcement learning to equip LVLMs with explicit multimodal reasoning abilities, improving both interpretability and performance. Specifically, we first warm up the LVLMs with SFT combined with MP-CoT, enabling the model to better capture multimodal associations and develop basic reasoning skills. The warm-up phase provides a stronger starting point for reinforcement learning, amplifying its effectiveness. Then, the model performs On-Model Sampling to get a group of responses, followed by reinforcement learning based on sampled predictions. Notably, the entire training process is parameter-efficient. Our main contributions can be summarized as follows:

- We conduct the first exploration of LVLMs on the JMERE task, extending reasoning-augmented paradigms, such as CoT and RL from pure-text LLMs to the multimodal setting.
- To address issues such as single-path reasoning and weak multimodal association in existing reasoning-augmented paradigms, we propose JMERE-R1, a multimodal training approach that requires no costly human-annotated CoT data. It guides LVLMs to effectively capture multimodal associations and perform diverse, interpretable reasoning chains for quintuple extraction.
- Extensive experiments show that our method significantly improves LVLM performance on JMERE, enhances the diversity of multimodal reasoning paths, and boosts interpretability—offering a strong baseline and clear direction for future research.

2 Related Work

Joint Multimodal Entity-Relation Extraction. Early research (Yao et al., 2019; Xie et al., 2022; Gao et al., 2024) in information extraction primarily focused on the text modality, with relatively little exploration of other modalities. With the rise

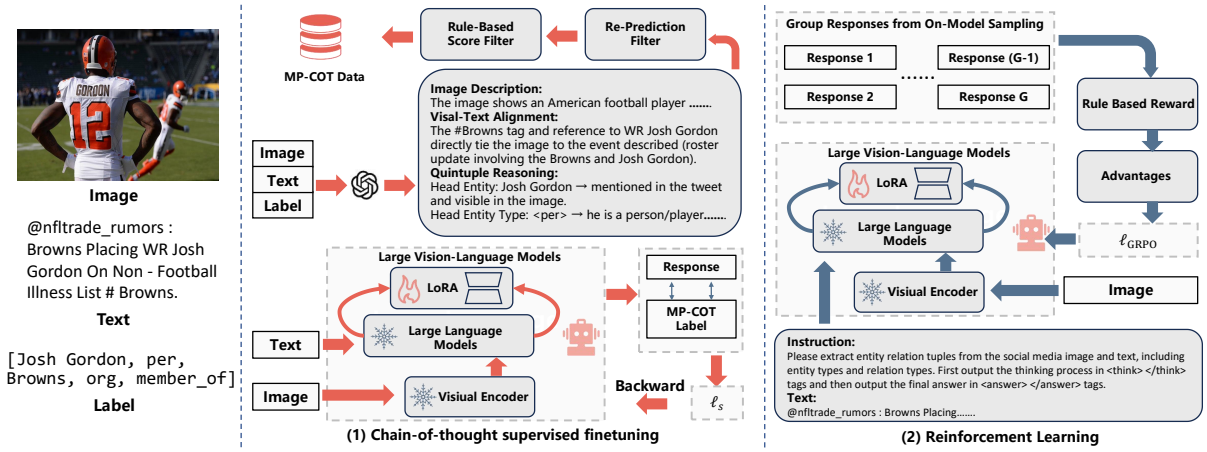

Figure 2: Overview of JMERE-R1. The training process consists of two stages. In the first stage, distilled MP-CoT data helps LVLMs acquire initial reasoning and format-following capabilities. In the second stage, reinforcement learning enable LVLMs to enhance their abilities through self-exploration of multiple potential reasoning paths.

of multimodal data on social media, MNER (Lu et al., 2018) and MRE (Zheng et al., 2021b,a) become key tasks in information extraction, aiming to leverage image data to enhance named entity recognition and relation extraction performance. Previous methods overlooked the interaction between these two tasks. Yuan et al. (Yuan et al., 2023) propose the JMERE task, which jointly performs MNER and MRE. Current multimodal information extraction methods (Xu et al., 2022; Cui et al., 2024; Wang et al., 2024) generally rely on additional image-text alignment modules and finetuning, including aligning entire images to vectors (Yu et al., 2020), aligning visual objects with textual counterparts (Wu et al., 2020; Zheng et al., 2020), and node alignment based on both text and visual graphs (Zhang et al., 2021). However, the performance of these methods is limited by the bottleneck of the alignment module's capability, making it difficult to model the complex cross-modal semantic relationships, which in turn prevents full exploitation of the model's potential. Furthermore, these methods rely on black-box discriminative decisions, which lack interpretability in their reasoning processes. This motivates us to explore a novel paradigm to address the JMERE task.

Reasoning enhanced Large Visual Language Models. LVLMs (Liu et al., 2023; Bai et al., 2025) are transformative in multimodal understanding and interaction. By seamlessly integrating visual perception with the natural language processing power of large language models, these models are redefining how AI interprets and understands complex information. LVLMs are widely used across various downstream tasks (Liu et al., 2024; Shao et al., 2024). Currently, LVLMs typically consist of three components: a visual encoder, a cross-modal projector, and a large language model (LLM). Using LVLMs for the JMERE task effectively leverages their rich pretraining knowledge and robust multimodal understanding capabilities. This drives us to explore how to better apply LVLMs to the JMERE task. Recently, RL methods (Liu et al., 2024; Shao et al., 2024) gain attention for enhancing the reasoning ability of LLMs. However, such work has primarily focused on text modalities, with few attempts in the multimodal space. Our work, starting from JMERE, proposes a training framework that encourages LVLMs to focus more on multimodal content and its associations during reasoning, thereby improving model performance.

3 Method

Figure 2 illustrates the overall architecture of JMERE-R1. First, a filtering mechanism is applied to vision-capable LLMs to obtain MP-CoT data. Then, this data is used with PEFT methods to perform SFT warm-up training on the LVLM. Finally, reinforcement learning is applied to the LVLM, enabling the model to self-explore and fully exploit the capabilities of both the vision and language modules for downstream tasks.

3.1 Problem Formulations

Given an input text sequence $D = \{d_1, d_2, ..., d_n\}$ and a corresponding image M, the task is to extract a set of quintuples: $y = \{(e_1, t_1, e_2, t_2, r)\}$ where each quintuple consists of two entities e_1 and e_2, their corresponding types t_1 and t_2, and a

relation type r between them. Our goal is to enable LVLMs to jointly understand text and image, generate the relevant quintuples. Let the input instruction be denoted as I, and the output token sequence generated by the model as $Y = f_{LLM}(I, D, M)$ where Y is the response produced by the LVLM. The generation probability of Y is defined as: $P(Y \mid I, D, M) = \prod_{t=1}^{l} P(y_t \mid I, D, M, y_{<t})$ where y_t denotes the t-th token in Y, and $y_{<t}$ represents all tokens generated before step t. Figure 2 illustrates an example from the JMERE task, where the model extracts the quintuple (Josh Gordon, per, Browns, org, member of)

3.2 CoT-SFT

MP-CoT Data. To facilitate the JMERE task, we propose a reasoning template named MP-CoT to guide the model in attending more effectively to visual content and cross-modal interactions. MP-CoT consists of three components: **Image Description**, which captures visual semantics; **Visual-Text Alignment**, which highlights associations between image and text; and **Quintuple Reasoning**, which supports logical inference for structured information extraction. We explicitly define the content of each component in the instruction prompts, and input the image, corresponding text, and ground-truth labels into a vision-capable GPT-4o[1] to obtain MP-CoT data. For each instance, we generate MP-CoT candidates five times, thereby enriching the candidates. The instruction templates for generating MP-CoT data can be found in Appendix D.

Re-Prediction Filter. To further improve the quality of the CoT data, we introduce a Re-Prediction Filter mechanism. Specifically, the image, text, and generated reasoning chain are fed back into the original model to produce a predicted set of quintuples. If the prediction aligns with the reference labels, the sample is retained.

Rule-Based Score Filter. For all samples that can perform re-prediction, we further design a scoring filtering mechanism that does not require training. Specifically, our scoring focuses on the accuracy of image descriptions, the capture of image-text associations, and the completeness of the reasoning process. We use rules for accurate descriptions and In-context learning methods to enable the model to score better. The detailed scoring criteria of the filter, the prompt templates used by the two filtering mechanisms, and the statistics of sample counts before and after filtering can be found in Appendix B.

[1]The version we use is gpt-4o-2024-11-20

Multimodal SFT. At this stage, we perform several epochs of fine-tuning on the LVLMs using MP-CoT data, enabling the model to acquire basic multimodal reasoning and quintuple extraction capabilities, and to generate coherent outputs. The input format is denoted as (X, E), where $X = \{I, D, M\}$ consists of the Instruction I, the accompanying text D, and the image M; E represents the reasoning data generated under a fixed MP-CoT paradigm. Let the set of reasoning data be denoted as P. To align with the ultimate goal of quintuplets extraction, and based on the token-level probability formulation described in the Problem Formulations section, we train the model using an autoregressive generation loss ℓ_s:

$$\ell_s = -\sum_{(X,E,L_r)\in P} \sum_{t=1}^{l_r} \log P_{\theta+\theta_r}(y_t | I, D, M, y_{<t}) \tag{1}$$

where l_r denotes the length of the quintuple extraction label sequence L_r. θ represents the parameters of the LVLMs, including both the vision and the LLM components, which are kept frozen during training. θ_r refers to the trainable parameters introduced by e Low-Rank Adaptation (LoRA).

3.3 Reinforcement Learning

We further enhance the LVLM's reasoning ability during self-exploration by applying the Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024) to the model. Specifically, we sample multiple responses from the model and assign reward values to each using a rule-based approach. Compared to using reward model (Schulman et al., 2017) to get reward score, this method is more computationally efficient and allows for direct supervision from gold-standard labels. The input prompt template can be found in Appendix E.

Accuracy-Based Rewards. We use the On-Model Sampling strategy to generate a set of G responses for each prompt. For a given response i, the corresponding gold label contains m quintuples, while the decoded output contains n predicted quintuples, among which c quintuples are correct. Inspired by standard evaluation metrics, we define

a basic quintuple accuracy reward r_q as:

$$r_i^q = \frac{2 \cdot \frac{c}{n} \cdot \frac{c}{m}}{\frac{c}{n} + \frac{c}{m}} = \frac{2c}{m+n} \tag{2}$$

Note that any response with formatting errors or invalid quintuple structure during LVLM self-exploration is assigned a reward of zero.

Partial Reward. Given that each JMERE instance typically includes only a small number of quintuples, we introduce an additional partial reward r_i^p to mitigate reward sparsity (Riedmiller et al., 2018; Trott et al., 2019). Let w denote the total number of correctly predicted components from $(e1, t1, e2, t2, r)$ across all quintuples. Let $\hat{Y}$ be the model output and Y the gold label. The partial reward is defined as:

$$r_i^p = \frac{w}{5n} + (1 - \frac{ED(\hat{Y}, Y)}{max(|\hat{Y}|, |Y|)}) \tag{3}$$

where ED denotes the edit distance between the predicted and gold responses. This fine-grained reward provides useful learning signals, especially when full matches are rare.

Format Reward. Inspired by recent advances in reasoning enhancement method (Liu et al., 2024), to further guide the model toward structured reasoning and extraction, we introduce a format reward r_i^f. Specifically, the model's response must follow a predefined structure containing both *<think> ... </think>* and *<answer> ... </answer>* segments. This constraint aligns with the SFT training format and helps preserve response consistency during self-exploration:

$$r_i^f = \begin{cases} 1, & \textit{if valid format} \\ 0, & \textit{if invalid format} \end{cases} \tag{4}$$

The overall reward r_i for each response is computed as:

$$r_i = r_i^q + \alpha r_i^p + \beta r_i^f \tag{5}$$

where α and β are coefficient factor for the partial reward and the format reward. The advantage $\hat{A}_i$ for response i is then calculated as:

$$\hat{A}_i = \frac{r_i - mean(r)}{std(r)} \tag{6}$$

where $mean$ denotes the average, and std denotes the standard deviation, r denotes the set of rewards corresponding to the G generated responses. This introduces an intra-group competition mechanism, guiding the model toward responses with higher relative advantage. To optimize the policy, we adopt the Group Relative Policy Optimization (GRPO) objective. Given a query $q = \{I, D, M\}$, the loss function is defined as:

$$\ell_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \left[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})} \cdot \hat{A}_{i,t} - \gamma \cdot \mathbb{D}_{\text{KL}}\left(\pi_\theta \,\|\, \pi_{\text{ref}}\right)\right] \tag{7}$$

where π_θ, $\pi_{\theta_{old}}$ and π_{ref} represent the log-probabilities assigned to the correct tokens by the current policy model, the old policy model at the beginning of the training step, and the reference policy model obtained from SFT-warmed parameters, respectively. $|o_i|$ is the output length of the i-th response, and t indexes the tokens within that response. The KL divergence term is weighted by γ to ensure the updated policy stays close to the reference distribution.

It is worth noting that, compared to the original GRPO method, we apply a simplified version to reduce computational overhead. In our implementation, only a single update is performed per group of generated responses, and we omit the advantage clipping step (Shao et al., 2024; Schulman et al., 2017). Similar to SFT, the training in this stage is also parameter-efficient.

4 Experiments

4.1 Datasets

We conduct our experiments using the JMERE dataset (Yuan et al., 2023), which is composed of MNER (Lu et al., 2018) and MRE (Zheng et al., 2021a) and excludes samples from the original dataset that lack entity types or relationships between entities. To the best of our knowledge, this is currently the only publicly available dataset for the Joint Multimodal Entity-Relation Extraction (JMERE) task. The entity types include Person, Organization, Location, and Miscellaneous, while the relation types consist of the 22 categories defined in prior work (Zheng et al., 2021a). We report precision, recall, and F1 score for both quintuple extraction and entity extraction as evaluation metrics for model performance. A quintuple is considered correct only when the entity, entity type, and relation type perfectly match the ground truth label.

Model/Micro F1(%)		JMERE			MNER		
		Precision	Recall	F1	Precision	Recall	F1
Pipline Method	OCSGA+MEGA	48.21	47.99	48.10	75.27	72.32	73.77
	AGBAN+MEGA	47.87	48.28	48.57	74.78	73.69	74.23
	UMGF+MEGA	49.28	50.76	50.01	75.02	76.77	75.88
Joint Method	OCSGA*	52.11	47.41	49.64	77.13	75.03	76.07
	AGBAN*	51.07	48.89	49.95	76.57	75.82	76.19
	UMGF*	52.76	50.22	51.45	77.51	76.67	77.22
	EEGA	58.26	52.61	55.29	78.27	**78.91**	78.59
SFT	LLaVa-1.5-7B	55.81	46.56	50.77	79.75	74.93	77.27
	Qwen2.5-VL-7B	56.88	47.81	51.95	**81.12**	76.37	78.67
CoT-SFT	LLaVa-1.5-7B	50.81	39.22	44.27	77.70	71.95	74.71
	Qwen2.5-VL-7B	50.54	43.59	46.77	76.48	73.10	74.75
Only-RL	LLaVa-1.5-7B	39.05	34.84	36.83	70.43	68.40	69.40
	Qwen2.5-VL-7B	40.11	34.53	37.11	73.33	70.51	71.89
JMERE-R1	LLaVa-1.5-7B	57.07	52.34	54.60	80.12	77.81	78.95
	Qwen2.5-VL-7B	**58.99**	**54.84**	**56.84**	80.14	78.29	**79.20**

Table 1: Results on the JMERE benchmark. The scores of existing discriminative methods are from the previous paper (Yuan et al., 2023). All LVLM-related methods are new experiments. AGBAN* refers to using the word-pair relation tagging in the AGBAN model. The metric values of the best-performing methods are highlighted in bold.

The detailed description of our experimental setup can be found in Appendix A.

4.2 Baselines

Existing Method. Based on previous studies (Yuan et al., 2023), we combine existing MNER and MRE methods into a pipeline as our strong JMERE baseline, as shown in Table 1. These methods include OCSGA (Wu et al., 2020), AGBAN (Zheng et al., 2020), and UMGF (Zhang et al., 2021) for entity and corresponding type extraction, as well as MEGA (Zheng et al., 2021a) for entity-relation extraction. In addition, we apply word-pair relation tagging to the above baseline models for joint extraction, such as OCSGA*, AGBAN*, UMGF*, and EEGA (Yuan et al., 2023).

LVLMs. We explore the performance of LVLMs on the JMERE task under various conditions, including zero-shot, traditional SFT paradigm, CoT-SFT, direct RL, and LVLMs trained using the JMERE-R1 paradigm. For the base models, we select the most advanced LVLMs currently available on public benchmark (Tang et al., 2024), including Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and LLaVa-1.5-7B (Liu et al., 2023).

4.3 Main Rsults

The experimental results on the JMERE dataset are shown in Table 1. We can find that: (1) Taking Qwen-2.5-VL as an example, the proposed JMERE-R1 framework outperforms the plain SFT paradigm by **0.53** F1 in MNER and **4.89** F1 in JMERE. Compared to CoT-SFT, the improvements are **4.45** F1 in MNER and **10.07** F1 in JMERE. When compared to Only-RL, the MNER score improves by **7.31** F1 and the JMERE score by **19.73** F1. Relative to existing methods, we achieve a **0.61** F1 improvement in MNER and a **1.55** F1 improvement in JMERE. These results show that the JMERE-R1 training framework significantly enhances the performance of LVLMs on the JMERE task, with similar results across different base models, demonstrating robustness. (2) Our method shows limited improvement in the NER metric. We attribute this to the fact that the reward function in the RL stage focuses on the overall performance of the quintuples rather than named entity recognition. The improvement in MNER is a side effect of the increased accuracy of the quintuple. (3) Although we improve the quality of MP-CoT data with a dual-filtering mechanism and use this data to help the model gain reasoning abilities, models trained with only CoT-SFT still do not outperform the traditional SFT paradigm. This phenomenon results from the increased length of answers caused by the inclusion of CoT data, making the SFT training objectives no longer focus solely on the accuracy of the quintuple tokens. Moreover, the soft labels from distillation data are not entirely reliable, which further affects the performance. (4)

Model/(%)	Precision	Recall	F1
JMERE-R1	**58.99**	**54.84**	**56.84**
w/o Partial Reward	56.86	53.12	54.93
w/o Rule-Based Score Filter	56.58	51.72	54.04
w/o Re-Prediction Filter	55.86	50.62	53.11
w/o All Filter	55.08	48.28	51.46
CoT-SFT	50.54	43.59	46.77
w/o Rule-Based Score Filter	47.02	41.88	44.30
w/o Re-Prediction Filter	46.67	40.47	43.35
w/o All Filter	44.44	39.37	41.76

Table 2: Ablation experiment on the JMERE test set, with the base LVLM Qwen2.5-VL-7B-Instruct and JMERE F1 metric.

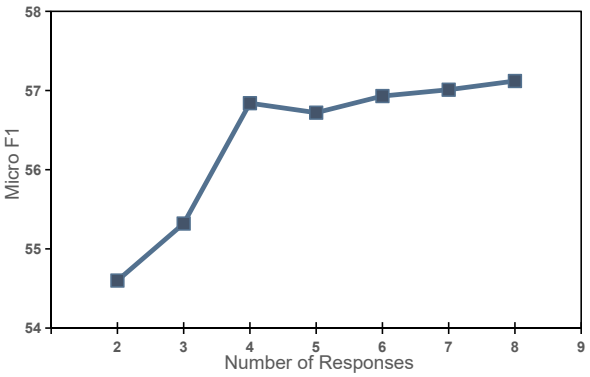

Figure 3: The Impact of Number of Generated Responses

When directly training LVLMs with RL, the initial model's understanding of the JMERE task is insufficient, especially in terms of task structure and the types involved. This leads to poor performance during the On-Model Sampling phase, with insufficient positive rewards. As a result, the model performs poorly under the Only-RL paradigm. This highlights the importance of CoT-SFT as a warm-up step.

4.4 Ablation Study

We perform ablation studies on the proposed JMERE-R1 framework using the test set to verify the effectiveness of each component in improving the model's performance. The experimental results are shown in Table 2. Below is a detailed analysis of each component:

w/o Partial Reward. We remove the Partial Reward and use only the Accuracy-Based Reward to calculate the advantage in reinforcement learning. The model's performance shows a noticeable decrease, demonstrating that the partial reward, which combines the quintuple components and character accuracy, helps reduce the impact of sparse rewards during training.

w/o Filter. We conduct separate and combined ablation studies on the two MP-CoT filtering mechanisms. We report results after both the CoT-SFT phase and the two-stage training phase. The results show that both the Re-Prediction Filter and the Rule-Based Score Filter contribute to improvements in the quality of CoT data. Additionally, the final model performance benefits from a better starting point at the end of the CoT-SFT phase.

4.5 Analysis and Discussion

The Number of Generated Responses. We explore the impact of the number of generated responses, G, on model performance during reinforcement learning training, with results shown in Figure 3. The experiment demonstrates that as the number of generated responses increases, the model's performance progressively improves. This suggests that in GRPO training, increasing the number of sampled responses helps the JMERE model more accurately identify the correct reasoning path among multiple possible options, thus assigning more appropriate advantages to each response. Further analysis reveals that as G increases, while the model's potential is not fully exploited, the performance gains begin to show diminishing returns as computational costs rise. This indicates that although increasing the number of responses can enhance model performance, excessively high sampling rates may lead to increased computational overhead, with performance improvements eventually leveling off. Therefore, in practical applications, it is crucial to adjust G appropriately to balance performance with computational efficiency.

Case Study To visually demonstrate the effectiveness of our proposed training framework, we present the responses generated by the Qwen-2.5-VL-Instruct model trained under different frameworks in Figure 4. Under the direct SFT paradigm, the model's training primarily focuses on fitting the existing data distribution, lacking a deep understanding of the interaction between image and text information. As a result, the model predicts incorrect quintuple. In the CoT-SFT paradigm, due to the incomplete reliability of MP-CoT data and the dispersion of the training process, the model erroneously identifies *Titan* as a player training on the field, leading to the prediction of incorrect quintuple. Compared to the SFT paradigm, CoT-SFT's visualization of the reasoning process allows us to quickly identify issues in the model and pro-

Image

RT@jwyattsports : .@Jack_Conklin78 putting in work @Titans practice

Text

[@Jack_Conklin78, per, @Titans, org, member_of]

Label

SFT

[["jwyattsports", "per", "Titans", "per", "Peer"]]

COT-SFT

<think> 1.**Description of the Image Content**: .... 2.**Connection Between the Text and the Image**: The image and text describe a training scene involving a person. 3. **Reasoning Process for Extracting the 5-Tuple:** Head Entity: "@Jack_Conklin78" (person).Tail Entity: "@Titans" (person), referring to an individual player as part of the team. Relation Type: "Peer", since both are within the same domain (the player's association with the team as part of the same group or practice).</think>
<answer> [["@Jack_Conklin78", "per", "@Titans", "per", "Peer"]] </answer>

RL

<think>\nThe image shows a group of individuals on a sports field, likely participating in a practice session. The individuals are wearing athletic gear, and some are holding protective equipment. The text \"RT @jwyattsports : .@Jack_Conklin78 putting in work @Titans practice\" indicates that the post is a retweet about a Titans player named Jack Conklin working during a practice session. The entity types and relation types can be inferred from the context provided by the text. </think>
<answer> [["Jack_Conklin78", "per", "Titans", "org", "Participation in practice"]] </answer>"

JMERE-R1

<think> 1.**Description of the Image Content**: The image depicts a football practice session with players in uniform, wearing helmets and practicing on a grassy field. Coaches and staff are also present, indicating an organized training environment. 2.**Connection Between the Text and the Image**: The text mentions @Jack_Conklin78 (likely a player associated with the Tennessee Titans, "@Titans") putting in work during practice. The image visually supports this claim by showing players actively practicing, aligning with the idea of training or work during practice. 3. **Reasoning Process for Extracting the 5-Tuple:** Head Entity: "@Jack_Conklin78" is identified as the primary subject mentioned in the text. Head Entity Type: "per" (person) because it refers to an individual player. Tail Entity: "@Titans" is mentioned in the text and is recognized as the organization associated with "@Jack_Conklin78." Tail Entity Type: "org" (organization) because it refers to the Tennessee Titans football team. Relation Type: "member_of" because the text implies that "@Jack_Conklin78" is a member of the "@Titans" organization and participating in their practice.</think>
<answer> [["@Jack_Conklin78", "per", "@Titans", "org", "member_of"]] </answer>

Figure 4: Case Study on JMERE: Responses of Qwen2.5-VL-7B under Different Thinking Enhancement Paradigms

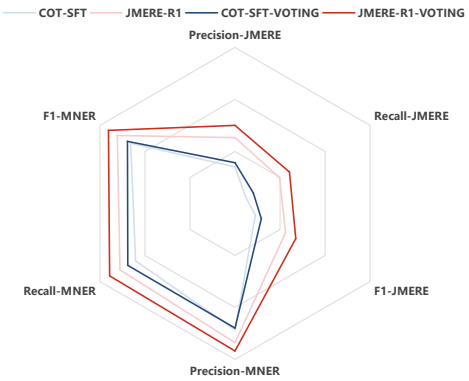

Figure 5: The performance variation of the model after Voting. The center point represents a value of 40% F1, while the outermost layer represents a value of 80% F1.

vides clear directions for optimization. In the RL paradigm, due to the model's lack of basic understanding of relationship categories, it fails to effectively exclude irrelevant relationship types during self-exploration. Despite having a logically correct reasoning process, the model still predicts an incorrect relationship type, such as *Participation in practice*. In the JMERE-R1 paradigm, with correct foundational knowledge and comprehensive self-exploration, the model is able to make logically sound inferences and accurately identify the correct quintuple.

Expansion of Reasoning Paths. To verify that the JMERE-R1 framework can expand the model's reasoning paths and enable flexible reasoning, we conducted multi-path reasoning experiments, following previous studies (Wang et al., 2023; Uesato et al., 2022; Trung et al., 2024). Specifically, for each test sample, we sampled 100 reasoning processes and corresponding answers, and integrated all results using majority voting. The experimental results are shown in Figure 5. As can be seen, compared to the CoT-SFT paradigm, JMERE-R1 achieves a significant performance boost after inheritance, reaching the best performance. This demonstrates the effectiveness of incorporating reinforcement learning and allowing the model to explore on its own, which expands the model's reasoning paths and pushes the performance limits of the model. By self-exploration, the model is able to select the optimal solution from multiple reasoning directions, further enhancing its reasoning capabilities and ability to handle complex tasks. This also indicates that enhancing the model's reasoning flexibility and diversity has a positive impact on improving its overall performance.

5 Conclusion

In this work, we propose a framework, JMERE-R1, for training LVLMs to better perform JMERE. First, we fine-tune LVLMs using distilled and filtered MP-CoT data to equip them with basic reasoning and quintuple extraction abilities. Then, we train the model with reinforcement learning by combining accuracy-based and partial reward functions, guiding the model to enhance its reasoning abilities through self-exploration. The partial reward function integrates the accuracy of the quintuples with character-level precision, helping to mitigate sparse reward issues and preventing reasoning path collapse. Experimental results and further analysis show that the model trained with our framework significantly outperforms existing methods on the public JMERE benchmark. JMERE-R1 effectively enhances the reasoning capabilities of LVLMs and broadens the reasoning paths, improve both the precision and diversity of the model's reasoning.

Limitations

The reward functions used during the reinforcement learning phase—including partial and format-based rewards—aim to guide the model's reasoning process. However, they often fail to fully capture the complexity of reasoning paths, leading to subpar performance in certain edge cases. The problem of reward sparsity still affects the model's accuracy. We design more fine-grained reward strategies to improve performance. Although we adopt parameter-efficient tuning and the GRPO training scheme to reduce computational cost, reinforcement learning with multi-response sampling still introduces considerable overhead. We further optimize the computational efficiency of our method in ongoing work.

References

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*.

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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Chufan Gao, Xuan Wang, and Jimeng Sun. 2024. Ttmre: Memory-augmented document-level relation extraction. *arXiv preprint arXiv:2406.05906*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

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

Martin Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas Degrave, Tom Wiele, Vlad Mnih, Nicolas Heess, and Jost Tobias Springenberg. 2018. Learning by playing solving sparse reward tasks from scratch. In *International conference on machine learning*, pages 4344–4353. PMLR.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, and 1 others. 2024. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*.

Alexander Trott, Stephan Zheng, Caiming Xiong, and Richard Socher. 2019. Keeping your distance: Solving sparse reward tasks using self-balancing shaped rewards. *Advances in Neural Information Processing Systems*, 32.

Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. Reft: Reasoning with reinforced fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7601–7614.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves

chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 257–268.

Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022. Maf: a general matching and alignment framework for multimodal named entity recognition. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 1215–1223.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. 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.

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

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.

Changmeng Zheng, Zhiwei Wu, Tao Wang, Yi Cai, and Qing Li. 2020. Object-aware multimodal named entity recognition in social media posts with adversarial learning. *IEEE Transactions on Multimedia*, 23:2520–2532.

A Implement Details

In our experiments, we use 4 A40-48GB GPUs for training and adopt DeepSpeed's Zero-2 stage optimization. For the SFT and CoT-SFT baseline models, we train the models for 20 epochs with a learning rate of 2e-5 and a warm-up ratio of 6%. We select the best-performing checkpoint on the validation set. This training duration is sufficient for the SFT model to converge. During training, the batch size per GPU is set to 2, resulting in a total batch size of 8. In the CoT-SFT warm-up phase, we train the model for 5 epochs with a learning rate of 2e-5. In the reinforcement learning phase and the RL baseline, we train the model for 30 epochs with a learning rate of 2e-6, and we select the best checkpoint on the validation set. The weight coefficient for partial reward, α, is set to 0.2, the weight coefficient for format reward, β, is set to 0.1, and the KL coefficient, λ, is set to 0.01. During testing, we uniformly set the generation temperature to 0.8, with a maximum generation length of 2048. In the Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning, we set the rank to 128 and the merging ratio to 64.

B Filter Details

In this section, we provide a detailed presentation of the prompt used by the dual-filtering mechanism. Figure 8 illustrates the prompt used by the Re-Prediction Filter, which inputs both the original data and the generated reasoning process into the model, and the generated results are used to filter out data that does not match the labels. Figure 9 presents the prompt used by the Rule-Based Score Filter, where we input the generated MP-CoT reasoning data and the original data into the model, along with the predefined discrete scoring rules. After scoring, we first filter out all samples with a score below 0.7. If a sample has multiple generated reasoning data, we select the one with the highest score. If there are multiple reasoning data with the same highest score, one is randomly selected.

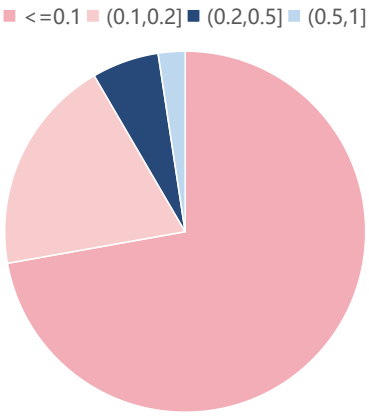

Figure 6: The error between GPT-4o and human scores.

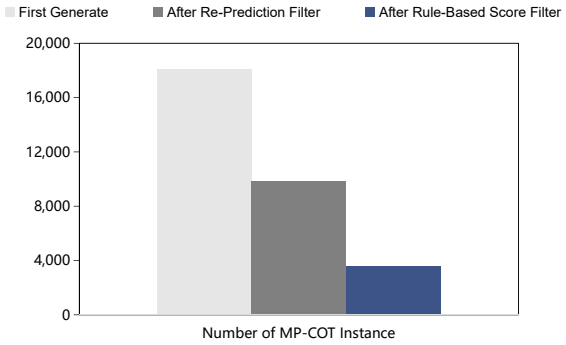

Figure 7: The change in the number of MP-CoT samples after filtering at each stage.

Reliability of the Scoring Filter. We conduct a manual evaluation to verify the reliability of the Rule-Based Score Filter. Specifically, we randomly select 100 samples from the training set, resulting in a total of 500 MP-CoT instances. Ten annotators are divided into two groups, with each person scoring 200 responses based on predefined rules. For each response, we calculate the average of four human-provided scores and use it as the reference to compute the error range of GPT-4o's scores. The results, shown in Figure 6, indicate that the majority of GPT-4o's scores fall within an error margin of 0.1, demonstrating that its scoring is reliable under fixed discrete rules.

Number of MP-CoT Instance. Figure 7 shows the change in the number of MP-CoT samples before and after the two-stage filtering mechanism. The results demonstrate that the filtering mechanism effectively removes noisy samples and low-quality reasoning data. The final sample count is 3534, compared to the original training set size of 3618, indicating that the filtering process effectively preserves sample integrity while removing invalid data.

C Zero-Shot Test

We test the performance of GPT-4o and two base models in a zero-shot setting. The The test prompt is the same as the RL prompt, except that the fixed-format reasoning instruction is removed. The Results are shown in Table 3. Overall, these models do not perform well, further proving the necessity of training LVLMs specifically for the JMERE task. These test results also provide a strong benchmark for future LVLM-based methods.

Model/(%)	Precision	Recall	F1
ChatGPT4o	9.95	17.47	12.68
Qwen2.5-VL-7B	7.27	14.10	9.60
LLaVa-1.5-7B	7.41	13.30	9.52

Table 3: Ablation experiment on the JMERE test set, with the base LVLM Qwen2.5-VL-7B-Instruct and JMERE F1 metric.

D MP-CoT Distill Prompt

We present the prompt template for obtaining MP-CoT data in Figure 10. In the prompt, we provide the image, text, and labels. Additionally, we specify the three components that must be included in the MP-CoT.

E RL Prompt

In this section, we provide a detailed description of the prompt used during the reinforcement learning phase of JMERE-R1, as shown in Figure 11. We input the text, image, and type ranges into the model and instruct it to output the reasoning results in a fixed format. The "Only-RL baseline" also uses the same set of prompts.

Image: 

Task: Based on the text, image, and reasoning process, provide the final five-tuple answer. The output format is: [Entity1, Entity Type1, Entity2, Entity Type2, Relation Type]

Text: @nfltrade_rumors : Browns Placing WR Josh Gordon On Non - Football Illness List # Browns.

Reasoning Process: <think>..... </think>

Figure 8: Prompt Template for Re-Prediction Filter

Image: 

Text: @nfltrade_rumors : Browns Placing WR Josh Gordon On Non - Football Illness List # Browns.

Task: Evaluate the reasoning process of a given answer and assign a score based on the following additive criteria. please only output the score
Scoring Rules:
Add +0.1 if the answer includes a description of an image.
Add +0.2 If the image description accurately reflects the image content.
Add +0.2 If the answer includes an analysis of the connection between the image and the accompanying text.
Add +0.3, if the reasoning process includes a detailed inference of the knowledge quintuple [Entity 1, Entity Type 1, Entity 2, Entity Type 2, Relation Type], including justification for entity extraction, rationale for determining each entity type, and the basis for relation type classification." It's not enough to simply infer where the knowledge quintuple is extracted from.
Add +0.2 If there is reflection that leads to overturning an incorrect conclusion, an additional 0.1 point may be awarded.

Reasoning Process: <think>..... </think> <answer> [["Josh Gordon", "per", "Browns", "org", "member_of"]] </answer>

Figure 9: Prompt Template for Rule-Based Score Filter

Image: 

Given an image and associated social media text, extract a 5-tuple in the form [head entity, head entity type, tail entity, tail entity type, relation type]. Provide the extraction process, including:

1. A description of the image content.

2. The connection between the text and the image.

3. A reasoning process for extracting the 5-tuple (keep it concise).

The text data is: {@nfltrade_rumors : Browns Placing WR Josh Gordon On Non - Football Illness List # Browns.}
The 5-tuple is:{[Josh Gordon, per, Browns, org, member of]}

Figure 10: Prompt Template for Obtaining MP-CoT Data

Image: 

Text: @nfltrade_rumors : Browns Placing WR Josh Gordon On Non - Football Illness List # Browns.

Please extract entity relation quadruples from the social media image and text, including entity types and relation types.

Range of entity types: <org> <misc> <per> <loc>

Range of relation types: <place_of_residence> <alumi> <neighbor> <awarded> <alternate_names> <part_of> <subsidiary> <siblings> <contain> <nationality> <race> <peer> <parent> <held_on> <locate_at> <religion> <member_of> <place_of_birth> <charges> <present_in> <couple>

The output format should be [[ent1, ent1 type, ent2, ent2 type, relation type]].

text: @nfltrade_rumors : Browns Placing WR Josh Gordon On Non - Football Illness List # Browns.

First output the thinking process in <think> </think> tags and then output the final answer in <answer> </answer> tags.

Figure 11: Prompt Template for reinforcement learning