Reasoning-Aware Relational Representation Learning for Open-Vocabulary Scene Graph Generation

Jiawei Xiao[1*], **Boya Wang**[1*], **Longtian Qiu**[1], **Shan Ning**[1,3], **Xuming He**[1,2†]
[1]ShanghaiTech University, Shanghai, China
[2]Shanghai Engineering Research Center of Intelligent Vision and Imaging
[3]Lingang Laboratory, Shanghai, China
{xiaojw2025, wangby2025, qiuilt, ningshan2022, hexm}@shanghaitech.edu.cn

Abstract

Open-Vocabulary Scene Graph Generation (OV-SGG) requires models to recognize visual relationships beyond the training vocabulary, yet existing methods often rely on dataset-specific object–predicate co-occurrence patterns. Under long-tailed distributions and noisy object localization, such reliance leads to biased, entangled relation representations, limiting both rare-relation recognition and generalization to unseen predicates. To address these challenges, we propose RaRe, a reasoning-aware relational representation learning framework that transforms MLLM-generated chain-of-thought (CoT) rationales into structured relation embeddings. Rather than using CoT only as an intermediate reasoning trace for final prediction, RaRe directly optimizes the rationale-derived representation space to improve inter-relation discriminability. Specifically, RaRe first aligns rationale embeddings with textualized relation descriptions through self-supervised contrastive learning, and then refines rationale generation with a specificity-oriented reinforcement learning objective that encourages semantically object-pair-specific relational evidence. Experiments on VG150 and PSG demonstrate improved category-balanced recognition over existing baselines, while the additional gains from learning vary across datasets and proposal settings. Code available at: https://github.com/dibur03/RaRe.

1 Introduction

Scene Graph Generation (SGG) represents an image as a structured graph of objects and their pairwise relationships, providing a compact abstraction that supports structured reasoning over visual content [31]. This representation has been widely adopted in vision-language tasks such as visual question answering [27, 12], image captioning [7, 37], and image retrieval [13]. While early SGG methods typically operated under a closed-set assumption with a fixed predicate vocabulary, recent work has shifted toward Open-Vocabulary Scene Graph Generation (OV-SGG), which enables models to generalize beyond the training vocabulary by capturing transferable relational semantics over novel object compositions and unseen predicates.

Existing OV-SGG approaches can be broadly categorized into two paradigms. *Embedding-matching methods* [11, 38, 4, 19] leverage vision-language alignment to transfer open-vocabulary recognition capabilities to relation prediction. In contrast, *generative methods* [22, 44, 34] cast the task as sequence generation, directly producing relational descriptions in natural language. Despite their progress, both paradigms exhibit several persistent limitations. First, generalization to unseen predicates remains limited, with performance often dropping significantly compared to seen relations,

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

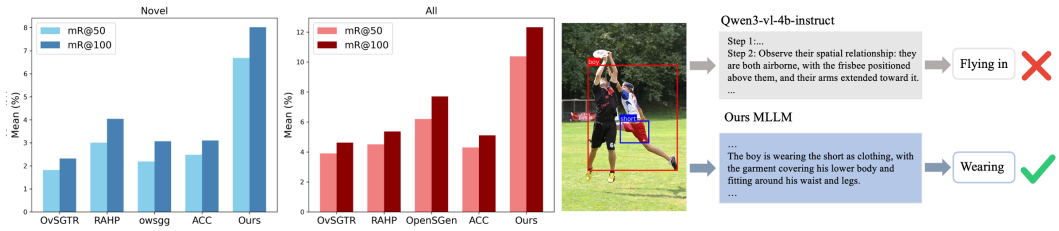


Figure 1: Left: our method achieves higher mR@50/100 on both novel and all predicate categories compared with existing OV-SGG methods on SGDet. Right: our method generates more object-pair-specific CoT relational evidence, leading to a better relation prediction than the baseline MLLM.

suggesting that current models tend to rely on dataset-specific object–predicate co-occurrence patterns rather than learning transferable relational semantics. Second, robustness under the realistic Scene Graph Detection (SGDet) setting is still inadequate. Relation prediction is highly sensitive to object localization errors, as inaccurate object candidate pairs can degrade relational reasoning. Third, long-tailed data distributions pose a more fundamental challenge in OV-SGG, as training is dominated by frequent object–predicate co-occurrence patterns, leading to biased and entangled representations that collapse toward frequent predicates, limiting both tail discrimination among seen classes and generalization to unseen ones. Consequently, models exhibit degraded performance on rare relations and limited transferability beyond the training distribution, as reflected in the persistent gap between overall recall (R@K) and mean recall (mR@K).

To address these challenges, we revisit OV-SGG from the perspective of relational representation learning. Rather than focusing on how to better utilize MLLMs for prediction, we treat them as a source of structured relational signals. Specifically, we propose a *reasoning-aware* relational representation learning framework that extracts instance-level rationales from MLLMs—such as spatial configurations, object attributes, and action states—and encodes them into structured embeddings for downstream relation prediction. By grounding relation representations in these compositional cues, our approach mitigates reliance on dataset-specific object–predicate co-occurrence patterns and promotes more transferable relational semantics. A key challenge, however, lies in the nature of these rationale signals: since MLLMs are trained with generative objectives that prioritize linguistic plausibility, the resulting representations are not optimized for inter-class discriminability. This often leads to entangled embeddings for semantically similar relations and exacerbates representation collapse under long-tailed distributions. To address this, our framework combines contrastive alignment of rationale–prediction pairs with specificity-driven policy optimization to learn relation-discriminative representations for open-vocabulary retrieval.

Concretely, we adopt a two-stage learning strategy. In the first stage, we employ self-supervised alignment to explicitly match MLLM-generated chain-of-thought (CoT) rationales with textualized relation descriptions, constructing a unified semantic space that captures fine-grained relational distinctions. In the second stage, we introduce reinforcement learning to refine the rationale generation process, encouraging the model to produce CoT outputs with clearer semantic separation across relation categories. Through this two-stage optimization process, the framework progressively learns more stable, discriminative, and transferable relational representations.

Experiments on Visual Genome (VG150) demonstrate that our approach improves OV-SGG performance across several key dimensions. In particular, the model achieves strong generalization to novel predicates (22.49 Novel mR@50) while maintaining balanced recognition on base categories (25.53 mR@50) under the PredCls setting. Under SGDet, RaRe achieves 10.38 Total mR@50 on VG150 and 17.52 on PSG using detector-generated proposals. Ablations further distinguish the contributions of the underlying retrieval pipeline from the setting-dependent gains of representation learning and calibration.

Our main contributions are summarized as follows:

- We propose a reasoning-aware relational representation learning framework for OV-SGG that addresses challenges in novel relation generalization, robustness of predicate recognition under noisy localization, and long-tailed distributions.

- We introduce structured MLLM-generated rationales as instance-level semantic priors for relational representation, together with a two-stage training strategy that integrates self-supervised alignment and reinforcement learning to extract and refine discriminative relational cues.
- Experiments validate that our method achieves consistent and notable improvements over existing approaches across open-vocabulary generalization tasks, realistic noisy detection (SGDet) settings, and long-tailed evaluation metrics.

2 Related Work

2.1 Scene Graph Generation

Scene Graph Generation (SGG) aims to localize objects and predict pairwise visual relationships as subject–predicate–object triplets in images [16]. Most existing methods formulate predicate prediction as a classification problem over a predefined set of categories [33, 39]. However, real-world data exhibits a long-tailed distribution of predicates, where frequent relations dominate training and rare ones are underrepresented, leading to biased predictions [28]. Moreover, these methods are limited to a fixed set of predefined categories and do not address generalization to unseen relations.

Open-Vocabulary SGG (OV-SGG) [11] extends conventional SGG by requiring models to recognize predicates unseen during training. Embedding-based methods leverage vision-language models such as CLIP [25] to align visual relation features with textual semantic embeddings [41, 4], with recent works further enhancing textual representations or incorporating interaction-aware prompting strategies [38, 19, 23] to improve novel predicate recognition. Another line of work adopts generative formulations: early approaches rely on constrained decoding guided by vision-language alignment [11, 22], while more recent methods employ multimodal large language models (MLLMs) to directly produce scene graphs as structured natural language [44, 34, 15]. Beyond the open-vocabulary setting, LLM4SGG [14] uses LLMs to augment relational supervision, and Relation-R1 [18] introduces CoT reasoning with reinforcement learning to boost their performance under open-ended settings. Despite their progress, existing methods generally optimize for generation accuracy or embedding alignment, without explicitly enforcing semantic separability across predicates—particularly under long-tailed distributions and noisy detection conditions. Our work bridges the embedding-based and generative paradigms by constructing a discriminative relational embedding space and using it to drive the MLLM's generative policy toward instance-specific outputs.

2.2 Chain-of-Thought Reasoning in Multimodal LLMs

Chain-of-Thought (CoT) prompting [32] enables step-by-step reasoning in LLMs and has been extended to multimodal settings [42]. In visual relation understanding, Relation-R1 [18] applies CoT reasoning for SGG, demonstrating that explicit intermediate reasoning can well enhance prediction quality. While this approach utilizes reinforcement learning to optimize CoT paths, its reward is defined at the prediction level, primarily based on the correctness of final outputs. Consequently, CoT is leveraged to guide reasoning and prediction, but is not explicitly modeled within a representation space where its semantic structure can be directly optimized. In contrast, we treat CoT rationales as relational representations and learn an embedding space to optimize their semantic specificity.

2.3 Reinforcement Learning and Contrastive Learning for LLMs

Reinforcement learning (RL) has emerged as a powerful tool for aligning LLM generation with task-specific objectives. Reinforcement Learning from Human Feedback (RLHF) [24] instantiates this paradigm by deriving reward signals from human preferences. Group Relative Policy Optimization (GRPO) [26, 8] is a policy optimization method that has been widely adopted for LLM post-training, where advantages are estimated via group-relative reward normalization, removing the need for a learned value network. These methods have achieved success in mathematical reasoning and instruction following, but their reward designs typically rely on ground-truth verification or human annotation, limiting applicability to tasks where labels are scarce or ambiguous.

In parallel, contrastive and self-supervised representation learning [3, 10, 29, 30, 9] has proven effective at encouraging inter-class separability in embedding spaces, underpinning many vision-

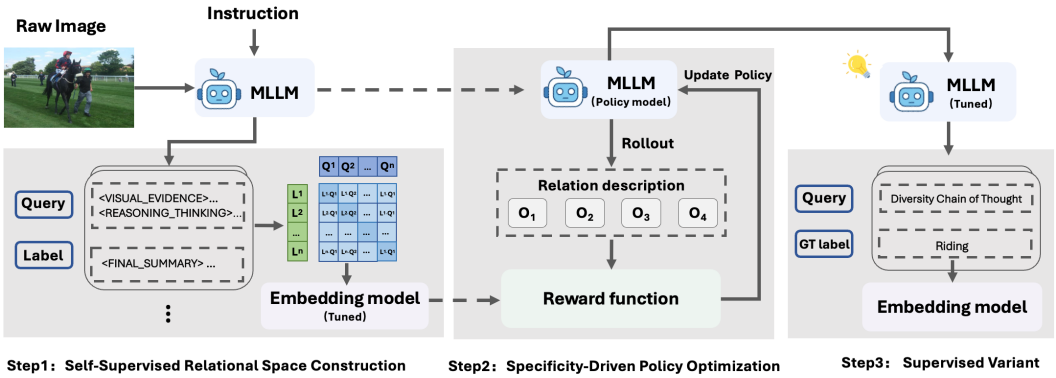


Figure 2: **Training pipeline of our method.** Step 1 we introduce Self-supervised training for the reward model. Step 2 we adopt GRPO for optimizing the MLLM. Step 3 we adopt supervised tuning for the final retrieval model.

language systems [25] and text embedding methods [6]. While effective for representation quality, these methods operate independently from the generative process of language models.

Our work unifies these two directions: we first learn a discriminative embedding space via self-supervised contrastive learning over CoT reasoning-prediction pairs, and then derive a label-free specificity reward from this space to optimize the MLLM's generation policy via GRPO—closing the loop between representation discriminability and generation quality.

3 Method

3.1 Preliminaries and Task Definition

Open-Vocabulary Scene Graph Generation. Given an image I, Scene Graph Generation (SGG) aims to produce a scene graph $\mathcal{G} = \{\mathcal{V}, \mathcal{R}\}$, where $\mathcal{V} = \{v_i\}_{i=1}^{N}$ is a set of entities with $v_i = (c_i^v, b_i)$ comprising a category label $c_i^v \in \mathcal{C}^v$ and a bounding box $b_i \in \mathbb{R}^4$, and $\mathcal{R} = \{r_{ij}\}_{i \neq j}$ is a set of relationships, each represented as a triplet $r_{ij} = \langle v_i, p_{ij}, v_j \rangle$ with predicate $p_{ij} \in \mathcal{C}^p$. In the *open-vocabulary* setting, $\mathcal{C}^p$ is split into base classes $\mathcal{C}^p_{\text{base}}$ and novel classes $\mathcal{C}^p_{\text{novel}}$ ($\mathcal{C}^p_{\text{base}} \cap \mathcal{C}^p_{\text{novel}} = \emptyset$). During training, only annotations from $\mathcal{C}^p_{\text{base}}$ are available, while at test time the model is required to generalize to both base and novel predicates.

MLLM-based Relation Prediction. We formulate predicate prediction as conditional text generation. Given an image I and a prompt $\mathbf{x}$ specifying the entity pair (v_i, v_j), the MLLM autoregressively generates a response $\mathbf{y} = (y_1, \ldots, y_T)$:

$$p_\theta(\mathbf{y} \mid I, \mathbf{x}) = \prod_{t=1}^{T} p_\theta(y_t \mid I, \mathbf{x}, \mathbf{y}_{<t}), \tag{1}$$

where θ denotes the MLLM parameters. The output consists of a chain-of-thought reasoning followed by a predicate prediction, denoted as $\mathbf{y} = [\mathbf{y}^{\text{cot}}; \mathbf{y}^{\text{pred}}]$.

Relation Scoring via Embedding Matching. To convert the MLLM's free-form response into structured predicate scores, we employ an embedding model f_ϕ that maps a text sequence to a normalized vector. Given the generated response y, we obtain its relation representation as $\mathbf{e} = f_\phi(y) \in \mathbb{R}^d$ with $\|\mathbf{e}\| = 1$. For each predicate category k, we construct a textual prototype $\mathbf{c}_k$ in the same embedding space. The relation score for category k is computed by cosine similarity:

$$s_k = \frac{\mathbf{e}^\top \mathbf{c}_k}{\|\mathbf{e}\| \|\mathbf{c}_k\|}. \tag{2}$$

This embedding-based scoring bridges CoT-style generation and structured relation prediction, and is later instantiated with multi-template predicate prototypes in Sec. 3.5.

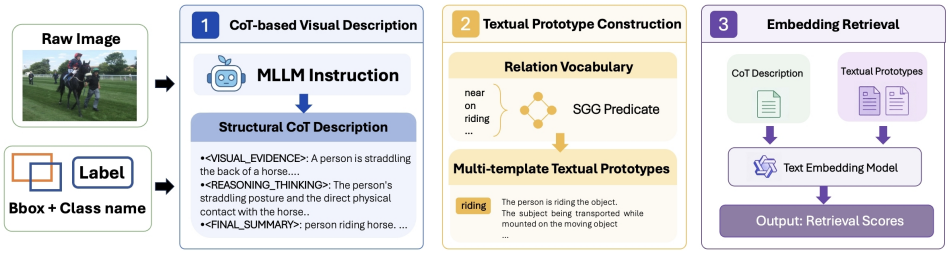


Figure 3: **Inference pipeline of our method.** (1)CoT-based Visual Description: The MLLM generates a structural CoT Description as the image's textual representation.(2)Textual Prototype Construction: Build up a Multi-template textual prototype for each SGG predicate.(3)Embedding Retrieval: We utilize an embedding model to compute the similarities between the CoT and the text prototypes to yield the final results.

3.2 Overview

We propose a two-stage framework for open-vocabulary scene graph generation that learns transferable and stable relational semantics beyond training distributions. As shown in Fig. 2, our framework consists of two core stages, followed by an optional supervised retrieval tuning step. In Stage 1, we learn a relational embedding space via contrastive learning by leveraging the intrinsic pairing between chain-of-thought (CoT) reasoning and relation predictions (Sec. 3.3). Before Stage 2, we perform cold-start supervised fine-tuning using responses generated by Qwen3-VL-30B-A3B-Instruct for 2,000 object-pair instances. In Stage 2, we optimize the MLLM via reinforcement learning, where rewards are derived from the relational embedding space learned in Stage 1 (Sec. 3.4). Overall, the embedding space learned in Stage 1 serves as a bridge between free-form reasoning and structured relation prediction, and provides guidance for optimizing the generation process in Stage 2.

3.3 Stage 1: Self-Supervised Relational Space Construction

To support generalization beyond seen predicates, we first aim to learn a transferable relational representation space that captures consistent semantics across diverse instances. A key challenge is that explicit annotations for fine-grained relations are limited in open-vocabulary settings.Our key observation is that the MLLM's own outputs already contain a naturally paired structure—each response couples a chain-of-thought reasoning $\mathbf{y}^{\text{cot}}$ with a relation prediction $\mathbf{y}^{\text{pred}}$—providing a self-supervisory signal for contrastive alignment without ground-truth predicate targets. Rather than treating predictions as discrete labels, we utilize this paired structure to learn representations grounded in reasoning, enabling the model to capture instance-level relational semantics while reducing reliance on dataset-specific co-occurrence patterns.

Data Collection. We first construct the training data for f_ϕ by performing inference with the MLLM over the training set. For each image I and entity pair (v_i, v_j), we query the MLLM with a unified relation-reasoning prompt that specifies the object categories and bounding boxes, and sample a response $y = [y^{\text{cot}}; y^{\text{pred}}]$. This yields a dataset $\mathcal{D} = \{(y_n^{\text{cot}}, y_n^{\text{pred}})\}_{n=1}^{|\mathcal{D}|}$ of chain-of-thought reasoning and predicted relation pairs. The full prompt template is provided in Appendix B.

Self-supervised Representation Learning. We train f_ϕ to align the CoT reasoning $\mathbf{y}^{\text{cot}}$ with its corresponding relation prediction $\mathbf{y}^{\text{pred}}$ via contrastive learning. Given a mini-batch of B pairs $\{(\mathbf{y}_i^{\text{cot}}, \mathbf{y}_i^{\text{pred}})\}_{i=1}^{B}$, we compute their embeddings:

$$\mathbf{e}_i^{\text{cot}} = f_\phi(\mathbf{y}_i^{\text{cot}}), \quad \mathbf{e}_i^{\text{pred}} = f_\phi(\mathbf{y}_i^{\text{pred}}). \tag{3}$$

The training objective is a symmetric InfoNCE loss. Taking the CoT-to-prediction direction as an example:

$$\mathcal{L}_{\text{c2p}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\text{sim}(\mathbf{e}_i^{\text{cot}}, \mathbf{e}_i^{\text{pred}})/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(\mathbf{e}_i^{\text{cot}}, \mathbf{e}_j^{\text{pred}})/\tau)}, \tag{4}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature hyperparameter. The prediction-to-CoT direction $\mathcal{L}_{\text{p2c}}$ is defined analogously. The overall objective is:

$$\mathcal{L}_{\text{stage1}} = \frac{1}{2}(\mathcal{L}_{\text{c2p}} + \mathcal{L}_{\text{p2c}}). \tag{5}$$

This objective aligns each CoT rationale with its corresponding predicted relation text, encouraging instance-level consistency in the learned representation space. Importantly, by relying on diverse CoT-prediction pairs rather than directly fitting base-class annotations, the learned space reduces dependence on annotated predicate frequencies and provides a transferable foundation for downstream alignment in Stage 2.

3.4 Stage 2: Specificity-Driven Policy Optimization

With the transferable representation space established by f_ϕ, we next aim to align the MLLM's generation process with this space, such that the produced reasoning reflects instance-specific relational semantics rather than generic or biased patterns. We adopt Group Relative Policy Optimization(GRPO) [26] and design a self-supervised reward grounded in the learned embedding space.

Training Objective. We present the training objective and highlight its key difference from vanilla GRPO [26]. Given an image I and a prompt $\mathbf{x}$ specifying an entity pair, the current policy π_θ generates a group of G responses $\{\mathbf{y}_i\}_{i=1}^{G}$, where each $\mathbf{y}_i = [\mathbf{y}_i^{\text{cot}}; \mathbf{y}_i^{\text{pred}}]$. A reward r_i is computed for each response, and the group-level advantage is estimated as $\hat{A}_i = (\boldsymbol{r_i} - \text{mean}(\{\boldsymbol{r_j}\}))/\text{std}(\{\boldsymbol{r_j}\})$. The training objective is:

$$\begin{aligned}
\mathcal{L}_{\text{stage2}}(\theta) &= -\mathbb{E}\left[\frac{\sum_{i=1}^{G}\sum_{t=1}^{T_i}\left(\tilde{u}_{i,t} - \beta\widehat{D}_{\text{KL},i,t}\right)}{\sum_{i=1}^{G} T_i}\right], \\
u_{i,t} &= \min\Big(\rho_{i,t}\hat{A}_i,\ \text{clip}(\rho_{i,t}, 0.8, 1.3)\hat{A}_i\Big), \\
\tilde{u}_{i,t} &= \begin{cases} \max(u_{i,t}, 3\hat{A}_i), & \hat{A}_i < 0, \\ u_{i,t}, & \hat{A}_i \geq 0. \end{cases}
\end{aligned} \tag{6}$$

Here, T_i is the number of valid response tokens and $\rho_{i,t} = \pi_\theta(y_{i,t} \mid I, \mathbf{x}, \mathbf{y}_{i,<t})/\pi_{\theta_{\text{old}}}(y_{i,t} \mid I, \mathbf{x}, \mathbf{y}_{i,<t})$. The expectation is over training samples and responses drawn from $\pi_{\theta_{\text{old}}}$. We average over valid response tokens and use a low-variance KL estimator relative to π_{ref} with $\beta = 0.01$. The response-level advantage $\hat{A}_i$ is computed from our specificity-based reward.

Representation-Aligned Reward. The reward encourages object-pair-specific responses without directly evaluating ground-truth predicate correctness. We compute this reward in the embedding space learned by f_ϕ in Stage 1. Concretely, consider a training batch consisting of M samples (entity pairs), each with G rollouts, yielding $M \times G$ responses in total. We encode each complete response $\mathbf{y}_i$ (the concatenation of CoT reasoning and relation prediction) into a normalized embedding:

$$\mathbf{e}_i = f_\phi(\mathbf{y}_i), \quad i = 1, \ldots, M \times G. \tag{7}$$

We sample each batch without repeated annotated subject–predicate–object category triplets to reduce false negatives. For a given rollout $\mathbf{y}_i$ belonging to sample m, we define its negative set $\mathcal{N}_i$ as all rollouts from *other* samples in the batch (i.e., rollouts from different entity pairs). The specificity reward is then computed as:

$$r_i^{\text{spec}} = 1 - \tau \cdot \left(\log \frac{1}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} \exp\big(\mathrm{sim}(\mathbf{e}_i, \mathbf{e}_k)/\tau\big)\right), \tag{8}$$

where τ is a temperature parameter and $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. The inner term is a LogMeanExp aggregation that provides a smooth approximation to the maximum similarity with negative samples. This formulation discourages generic responses shared across different entity

pairs by assigning lower rewards to responses with higher cross-pair similarity. It encourages object-pair-specific relational descriptions and supports learning more discriminative representations for downstream relation retrieval.

The specificity reward does not directly verify predicate correctness or visual grounding; a distinctive rationale may still be visually incorrect. In addition to the Representation-Aligned Reward, we include a format reward r_i^{fmt} that checks whether the response follows the prescribed output structure (e.g., producing a valid chain-of-thought followed by a relation prediction). The final reward is:

$$r_i = r_i^{\text{spec}} + \lambda\, r_i^{\text{fmt}}, \tag{9}$$

where λ controls the relative weight of format compliance.

3.5 Relation Prediction

At test time, we construct predicate prototypes for the full predicate vocabulary $\mathcal{C}^p = \mathcal{C}^p_{\text{base}} \cup \mathcal{C}^p_{\text{novel}}$ using multiple textual templates for each predicate. Specifically, for each relation category k, we encode a set of template descriptions $\{\mathbf{c}_{k,m}\}_{m=1}^{M}$ with the embedding model, where $\mathbf{c}_{k,m} = f_\phi(T_m(\text{name}_k))$. Given an image I and an entity pair (v_i, v_j), the optimized MLLM generates a single response $\mathbf{y} = [\mathbf{y}^{\text{cot}}; \mathbf{y}^{\text{pred}}]$ via greedy decoding. The complete response is mapped to the embedding space as $\mathbf{e} = f_\phi(\mathbf{y})$. For each predicate, its score is computed as the maximum similarity between the response embedding and all its template prototypes, and the final predicate is determined by:

$$p_{ij} = \arg\max_{k} \max_{m=1,\ldots,M} \text{sim}(\mathbf{e}, \mathbf{c}_{k,m}). \tag{10}$$

This procedure is applied to all candidate entity pairs to produce the final scene graph.We further apply pair-level score normalization and a lightweight subject-object prior calibration during inference, with details provided in Appendix C.

3.6 Supervised Variant

To align with the standard OV-SGG evaluation protocol, where base class annotations are available during training, we introduce a supervised extension. Starting from the embedding model f_ϕ trained in Stage 1, we utilize the semantic embeddings of the base class relation names as textual anchors, denoted as $\{\mathbf{a}_k\}_{k=1}^{|\mathcal{C}^p_{\text{base}}|}$, where each anchor is generated by encoding the predicate text through f_ϕ. We then fine-tune f_ϕ with an InfoNCE objective analogous to (Eq. (5)), aligning each MLLM-generated reasoning output embedding with its corresponding ground-truth relation name embedding while contrasting it against all other base class candidates. At inference time, the model serves as a relation extractor, determining the predicate by identifying the nearest neighbor in the embedding space. Since novel class embeddings are constructed identically from their respective relation names, this approach naturally extends the model's discriminative power to unseen relations without requiring any novel class annotations.

4 Experiments

4.1 Experiment Setup

Datasets We evaluate our method on Visual Genome (VG150) [16] and PSG [36]. For VG150, we follow the standard *OVR-SGG* protocol [4], we use the EPIC [38] split for PredCls and the OvS-GTR [4] split for SGDet. Both contain 35 base and 15 novel predicates, with different memberships (Appendix C.1). During training, only base-predicate annotations are available, while evaluation is conducted on both base and novel predicates to measure open-vocabulary generalization. We report results under two standard settings: Predicate Classification (PredCls) and Scene Graph Detection (SGDet) [33]. For PSG, we follow PGSG's predicate split [22] and evaluate under SGDet. In addition, we provide supplementary Human-Object Interaction (HOI) experiments on HICO-DET [2] in Appendix H to further examine the transferability of our framework across visual relationship understanding tasks.

Table 1: **OVR-SGG results under SGDet on the full VG150 test set.** All results use graph-constraint evaluation. † denotes our re-evaluation of ACC. ACC+RaRe uses the same detected boxes, object labels, and candidate pairs as ACC.

Model	Total (Relation)		Base (Relation)		Novel (Relation)	
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
OvSGTR[4]	20.46/23.86	3.91/4.62	26.14/30.16	4.81/5.60	13.45/16.19	1.82/2.32
OvSGTR+RAHP[23]	21.50/25.74	4.51/5.37	26.29/30.16	5.15/5.94	15.59/19.92	3.01/4.04
OwSGG(Qwen72b)[5]	-/-	-/-	-/-	-/-	2.19/3.06	2.19/3.07
OpenSGen[15]	18.0/20.5	6.2/7.7	-/-	-/-	15.7/17.9	-/-
ACC(Swin-B)[19]†	**24.04/28.50**	4.30/5.11	27.41/31.53	5.08/5.95	**19.15/23.72**	2.48/3.17
ACC+Ours	19.57/22.96	**10.38/12.32**	**28.81/33.10**	**11.96/14.17**	9.94/12.38	**6.69/8.02**

Table 2: **Experimental results of OVR-SGG on PredCls setting.**

Model	Total (Relation)		Base (Relation)		Novel (Relation)	
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
PGSG[22]	26.90/33.90	10.80/13.90	-	-	-	5.20/7.70
SGTR+[21]	36.32/41.51	13.30/17.60	40.50/46.67	19.57/24.96	0.00/0.00	0.00/0.00
SGTR+ + RAHP[23]	39.92/46.03	16.88/22.18	41.29/47.65	20.51/25.18	15.46/20.37	11.82/15.46
SVPR[11]	33.50/35.90	8.30/10.80	-/-	-/-	-/-	-/-
Epic[38]	-/-	16.50/21.80	28.30/31.10	-/-	13.90/18.30	-/-
PE-NET[43]	58.79/61.23	19.18/20.97	63.62/67.09	23.18/25.79	0.00/0.00	0.00/0.00
PE-NET+RAHP[23]	**64.70/69.11**	24.50/**28.25**	**65.15/70.54**	24.99/**30.19**	**20.79/29.00**	15.70/23.73
Ours	49.90/51.33	**24.62**/25.80	50.36/51.81	**25.53**/26.45	12.75/13.60	**22.49/24.27**

Evaluation metric We adopt mean Recall@K ($mR@K$) and Recall@K ($R@K$) as the primary evaluation metrics, where $K \in \{50, 100\}$. While $R@K$ measures overall triplet-level recall and is often dominated by frequent predicates, $mR@K$ computes recall independently for each predicate category and then averages over all classes. Therefore, $mR@K$ provides a more balanced evaluation of low-frequency, long-tailed, and unseen relations, which is particularly aligned with the objective of OV-SGG. We thus regard $mR@K$ as the main metric and report $R@K$ as a complementary measure for comprehensive comparison.

Implementation Details We adopt Qwen3-VL-4B-Instruct [1] as the backbone MLLM and Qwen3-VL-Embedding-2B [20] as the embedding model f_ϕ. The embedding model is fine-tuned with LoRA on 2 NVIDIA L20 GPUs. The MLLM is fully fine-tuned with GRPO on 4 NVIDIA L40 GPUs. For SGDet, we use the object detector from ACC [19] to provide object proposals, isolating improvements in predicate prediction. For VG150 SGDet, we use TP, PR, SSL, and RL without supervised embedding tuning (ST). More implementation details are provided in Appendix C.

4.2 Comparisons with State-of-the-Art OV-SGG Methods

Comprehensive comparison between our framework and state-of-the-art (SOTA) OV-SGG methods under the standard SGDet and PredCls setting are in Table1 and Table 2. The key findings and conclusions are summarized as follows:

Performance under Detection Settings As shown in Table 1, RaRe achieves 14.17% Base mR@100 and 8.02% Novel mR@100 on VG150 using detector-generated proposals. Under fixed ACC proposals, training-free MLLM retrieval already accounts for most of the Total mR gain over ACC; adding SSL, RL, and PR yields modest further gains while reducing Novel mR (Table 16). In contrast, the same-proposal PSG ablation shows improvements in both Total and Novel mR over training-free retrieval (Table 5), indicating that the benefits of learning and calibration vary across datasets and proposal settings.

Improvement on Novel Class Generalization We further examine novel predicate generalization under both SGDet and PredCls settings, as shown in Tables 1 and 2. Under SGDet on VG150, RaRe achieves 10.38/12.32 Total mR@50/100, with 14.17% Base mR@100 and 8.02% Novel mR@100, improving category-balanced recognition of unseen predicates in the end-to-end setting.

Under PredCls, where ground-truth object boxes are provided, closed-set SGG methods adapted to the open-vocabulary setting, such as PE-NET [43] and SGTR+ [21], show limited generalization to

Table 3: **OVR-SGG results under SGDet on PSG.** PGSG+RaRe uses the same object proposals and candidate pairs as PGSG.

Model	Total (Relation)		Novel (Relation)	
	R@50/100	mR@50/100	R@50/100	mR@50/100
SGTR+ [21]	14.20/18.20	6.40/8.40	–/–	0.00/0.00
PGSG [22]	18.00/20.20	13.50/16.40	–/–	7.40/11.30
OvSGTR [4]	17.76/19.50	–/–	6.93/8.08	–/–
ACC [19]	**20.01/21.71**	–/–	8.78/9.70	–/–
PGSG+RaRe	16.34/18.65	**17.52/20.59**	**9.11/10.34**	**14.76/16.10**

novel predicates. Compared with the strongest VLM-based baseline PE-NET+RAHP [23], which achieves 15.70% Novel mR@50, RaRe improves this metric to 22.49%, yielding an absolute gain of 6.79 percentage points. These results indicate that reasoning-aware relational representations help transfer relational semantics beyond base predicates, leading to stronger generalization to unseen relations.

Balanced Recognition across Long-Tailed Distributions As shown in Table 2, RaRe achieves stronger category-balanced performance despite a lower Base R@50 than PE-NET+RAHP, reaching 25.53% Base mR@50 and 24.62% Total mR@50. This indicates that our method improves recognition beyond dominant head predicates and better covers the long-tailed predicate distribution. The contrast between lower global R@K and higher mR@K suggests that RaRe shifts the prediction behavior from fitting frequent co-occurrence patterns toward more balanced relational generalization.

Results on PSG We evaluate RaRe under SGDet on PSG using PGSG's predicate split, retaining its detected boxes, object labels, and candidate pairs. As shown in Table 3, PGSG+RaRe achieves the best reported Total mR, Novel mR, and Novel R among the compared methods, although its Total R remains lower than PGSG.These results extend the gains in category-balanced and novel-predicate recognition to another dataset and upstream detector.

4.3 Ablation Study

We conduct ablation studies under the PredCls setting to validate each component of our framework, with results summarized in Table 4.

Effectiveness of Core Modules Compared with TP-only retrieval, RaRe-Core(row 7) improves Total and Novel mR@50/100 by 3.90/3.93 and 5.01/4.41 points, respectively, demonstrating the contribution of reasoning representation learning without PR or ST. Adding PR and ST further improves Base mR@50/100 from 21.81/23.14 to 25.53/26.45, while reducing Novel mR@50/100 from 27.11/28.46 to 22.49/24.27. These results distinguish the gains from the core learning stages from the base–novel trade-off introduced by downstream tuning and calibration. We report the full configuration as a compromise between base and novel recognition: it improves Base mR over RaRe-Core while retaining higher Novel mR than the SSL+ST variant.

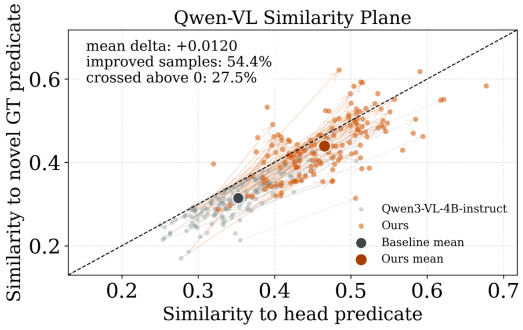


Figure 4: Distribution of CoT samples in the similarity plane.

Head Predicate Prediction Shift Analysis As shown in Figure 4, we project novel ground-truth (GT) samples onto a 2D similarity plane defined by $\mathrm{sim}(\mathrm{CoT}, \mathrm{head})$ and $\mathrm{sim}(\mathrm{CoT}, \mathrm{novel\ GT})$. The diagonal dashed line denotes equal similarity, so points above this line indicate CoT representations that are semantically closer to the novel predicate than to the head predicate. Compared with the

Table 4: **Ablation study of RaRe under PredCls on the full VG150 test set.** We evaluate the impact of different components: TP (predicate template for multi-template prototype construction), PR (prior calibration), SSL (self-supervised embedding learning in Stage 1), RL (GRPO in Stage 2), and ST (supervised embedding tuning). We report category-balanced mR@50/100 in the main paper.

#	TP	PR	RL	SSL	ST	Total mR@50/100	Base mR@50/100	Novel mR@50/100
1						19.42/20.25	17.94/19.02	22.88/23.13
2	✓					19.50/20.80	18.38/19.41	22.10/24.05
3	✓	✓				20.68/23.54	23.37/25.93	14.39/17.95
4	✓	✓	✓			20.80/23.25	21.39/23.61	19.41/22.41
5	✓	✓		✓		21.54/24.23	23.61/26.08	16.72/19.92
6	✓	✓		✓	✓	**27.29/29.14**	**33.11/34.61**	13.71/16.38
7	✓		✓	✓		23.40/24.73	21.81/23.14	**27.11/28.46**
8	✓	✓	✓	✓	✓	24.62/25.80	25.53/26.45	**2**2.49/24.27

Table 5: **Same-proposal SGDet ablation on PSG.** All variants use identical PGSG proposals and evaluation protocols, with TP and PR fixed. SSL, RL, and ST denote self-supervised embedding learning, policy optimization, and supervised embedding tuning, respectively.

#	RL	SSL	ST	Total mR@50/100	Novel mR@50/100
1				12.56/14.39	12.69/14.20
2		✓		14.02/16.68	14.25/15.88
3		✓	✓	15.12/17.04	9.47/10.49
4	✓			12.63/14.45	12.83/14.31
5	✓	✓		13.89/15.92	14.75/**16.34**
6	✓	✓	✓	**17.52/20.59**	**14.76**/16.10

baseline, our method exhibits a systematic shift toward higher similarity to the novel GT, together with a more dispersed distribution. This observation suggests that RL-aligned CoT representations reduce relative reliance on head predicates and better capture novel relational semantics.

Same-Proposal Ablation on PSG With PGSG proposals and TP/PR fixed, full RaRe improves Total mR@50/100 from 12.56/14.39 to 17.52/20.59 over training-free retrieval (Table 5). Adding RL to SSL+ST improves Novel mR@50/100 by 5.29/5.61 points, mitigating the decline introduced by supervised tuning. Full RaRe achieves the highest Total mR, while omitting ST yields slightly higher Novel mR@100.

5 Conclusion

In this work, we present RaRe, a reasoning-aware relational representation learning framework for open-vocabulary scene graph generation. Rather than treating MLLM outputs merely as final textual predictions, RaRe transforms chain-of-thought rationales into structured relation embeddings that capture object-pair-specific relational evidence. By aligning CoT rationales with relation semantics through self-supervised contrastive learning and further refining rationale generation with representation-guided policy optimization, our framework converts free-form MLLM reasoning into discriminative, retrieval-compatible relation representations. Experiments on VG150 and PSG demonstrate improved category-balanced predicate recognition, with ablations showing that the additional benefits of learning and calibration depend on the evaluation setting. These results suggest that rationale-derived relational representations provide an effective bridge between generative visual reasoning and structured open-vocabulary relation prediction.

Limitation This work has two main limitations. First, although our framework improves the utility of MLLM-generated rationales for relation prediction, it does not explicitly address all sources of visual hallucination in MLLMs, which may still cause fine-grained inaccuracies for small or occluded objects. Second, balancing category-level generalization and overall recall remains challenging, and the benefits of learning and calibration vary across proposal settings. Under fixed ACC proposals on VG150, novel-predicate recognition remains below the training-free retrieval baseline (Appendix J), possibly reflecting the mismatch between ground-truth training pairs and detector-generated test pairs. Future work could explore adaptive calibration and better adaptation to noisy proposals. The

framework may benefit image retrieval and assistive scene understanding, but incorrect or biased predictions can propagate to downstream applications.

6 Acknowledgments

This work was supported by the Shanghai Key Technology R&D Program No.25JC3200500, and the MoE Key Lab of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University).

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[4] Zuyao Chen, Jinlin Wu, Zhen Lei, Zhaoxiang Zhang, and Chang Wen Chen. Expanding scene graph boundaries: Fully open-vocabulary scene graph generation via visual-concept alignment and retention. In *European Conference on Computer Vision*, pages 108–124. Springer, 2024.

[5] Amartya Dutta, Kazi Sajeed Mehrab, Medha Sawhney, Abhilash Neog, Mridul Khurana, Sepideh Fatemi, Aanish Pradhan, M Maruf, Ismini Lourentzou, Arka Daw, et al. Open world scene graph generation using vision language models. *arXiv preprint arXiv:2506.08189*, 2025.

[6] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6894–6910, 2021.

[7] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10323–10332, 2019.

[8] Daya Guo, Dejian Yang, Haowei Zhang, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

[9] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in neural information processing systems*, 34:5000–5011, 2021.

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020.

[11] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Towards open-vocabulary scene graph generation with prompt-based finetuning. In *European conference on computer vision*, pages 56–73. Springer, 2022.

[12] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[13] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[14] Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. Llm4sgg: Large language models for weakly supervised scene graph generation, 2024.

[15] Zihan Kong and Haiwei Zhang. Opensgen: Fine-grained relation-aware prompt for open-vocabulary scene graph generation. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, ICMR '25, page 634–643, New York, NY, USA, 2025. Association for Computing Machinery.

[16] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[17] Ting Lei, Fabian Caba, Qingchao Chen, Hailin Jin, Yuxin Peng, and Yang Liu. Efficient adaptive human-object interaction detection with concept-guided memory. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6480–6490, 2023.

[18] Lin Li, Wei Chen, Jiahui Li, Kwang-Ting Cheng, and Long Chen. Relation-r1: Progressively cognitive chain-of-thought guided reinforcement learning for unified relation comprehension, 2025.

[19] Lin Li, Chuhan Zhang, Dong Zhang, Chong Sun, Chen Li, and Long Chen. Interaction-centric knowledge infusion and transfer for open-vocabulary scene graph generation. *arXiv preprint arXiv:2511.05935*, 2025.

[20] Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

[21] Rongjie Li, Songyang Zhang, and Xuming He. Sgtr+: End-to-end scene graph generation with transformer, 2024.

[22] Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models, 2024.

[23] Tao Liu, Rongjie Li, Chongyu Wang, and Xuming He. Relation-aware hierarchical prompt for open-vocabulary scene graph generation, 2025.

[24] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[25] 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. Learning transferable visual models from natural language supervision, 2021.

[26] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[27] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8376–8384, 2019.

[28] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[29] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[30] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.

[31] Yu-Siang Wang, Chenxi Liu, Xiaohui Zeng, and Alan Yuille. Scene graph parsing as dependency parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 397–407, 2018.

[32] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[33] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017.

[34] Mingjie Xu, Mengyang Wu, Yuzhi Zhao, Jason Chun Lok Li, and Weifeng Ou. Llava-spacesgg: Visual instruct tuning for open-vocabulary scene graph generation with enhanced spatial relations, 2024.

[35] Shiyu Xuan, Dongkai Wang, Zechao Li, and Jinhui Tang. Zero-shot hoi detection with mllm-based detector-agnostic interaction recognition. *arXiv preprint arXiv:2602.15124*, 2026.

[36] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *European conference on computer vision*, pages 178–196. Springer, 2022.

[37] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10685–10694, 2019.

[38] Qifan Yu, Juncheng Li, Yu Wu, Siliang Tang, Wei Ji, and Yueting Zhuang. Visually-prompted language model for fine-grained scene graph generation in an open world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21560–21571, 2023.

[39] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[40] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20072–20080. IEEE, 2022.

[41] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Learning to generate language-supervised and open-vocabulary scene graph using pre-trained visual-semantic space. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2915–2924, 2023.

[42] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024.

[43] Chaofan Zheng, Xinyu Lyu, Lianli Gao, Bo Dai, and Jingkuan Song. Prototype-based embedding network for scene graph generation, 2023.

[44] Zijian Zhou, Zheng Zhu, Holger Caesar, and Miaojing Shi. Openpsg: Open-set panoptic scene graph generation via large multimodal models, 2024.

A Overview of Appendix

In this supplementary document, we provide additional details and experimental results to enhance understanding and insights into our proposed RaRe. This supplementary document is organized as follows:

- B: Instruct Prompt for Inference
- C: Additional Implementation Details
- D: Zero-Shot PredCls Performance of Different MLLM Backbones
- E: Sensitivity analysis
- F: Comparison with Alternative MLLM Optimization Strategies
- G: SGCls Setting Results
- H: Cross-Task Transfer: Training-Free Zero-Shot HOI
- I: Additional Analysis on Representation Properties
- J: Additional Controls and Inference Cost
- K: Novel-Predicate Coverage under SGDet
- L: More Qualitative Comparison Results

B Instruct Prompt for Inference

This section shows the unified instruct prompt we use during the inference stage, which is designed to guide the MLLM to generate relation descriptions with chain-of-thought (CoT) reasoning based on the input entity pairs.

```
In this image, there are two objects:
- <|object_ref_start|>{subject_class}<|object_ref_end|> at
↪  <|box_start|>{subject_bbox}<|box_end|>
- <|object_ref_start|>{object_class}<|object_ref_end|> at
↪  <|box_start|>{object_bbox}<|box_end|>

Based on the image, please analyze the predicate relationship.
IMPORTANT: You must strictly follow the format below using the specific headers
↪  enclosed in brackets. Do not output any conversational filler.

<VISUAL_EVIDENCE>
Describe the spatial layout and visual details. Focus primarily on the two objects
↪  mentioned above.</VISUAL_EVIDENCE>
<REASONING_THINKING>
Provide step-by-step logical reasoning. Connect the visual evidence to the most
↪  likely relationship. Explain the interaction mechanism. Fully consider the
↪  relationships between objects.</REASONING_THINKING>
<FINAL_SUMMARY>
State the final relationship predicate in a 'subject predicate object' format and
↪  a brief summary sentence describing the interaction.</FINAL_SUMMARY>
```

C Additional Implementation Details

C.1 Predicate Splits.

For VG150, both settings divide the 50 predicates into 35 base and 15 novel categories, but use different splits. Under PredCls, we follow EPIC [38] and use *across, along, and, eating, flying in, growing on, laying on, lying on, mounted on, on back of, painted on, part of, says, using, walking in* as novel predicates. Under SGDet, we follow OvSGTR [4] and use *belonging to, part of, riding, walking in, in, of, painted on, playing, for, walking on, says, attached to, eating, on, wears* as novel predicates. The remaining 35 predicates form the base vocabulary in each setting. Training, prior construction, and evaluation use the corresponding split.

C.2 Self-supervised Training on Embedding

To provide a fine-grained and semantically-aware reward signal for Group Relative Policy Optimization (GRPO), we first train a specialized embedding model via self-supervised contrastive learning. This model serves as the reward encoder for evaluating the quality of generated Chain-of-Thought (CoT) reasoning in the subsequent GRPO stage.

Training Objective and Data. We construct training pairs using only base-set samples, where CoT reasoning is generated by a pretrained Qwen3-VL-4B-Instruct model [1] (official release). Given an image and an object pair, the model produces a structured CoT as shown in B, the <VISUAL_EVIDENCE> and <REASONING_THINKING> form the *query* r_i, while the <FINAL_SUMMARY> is the corresponding *label* s_i. The objective is to align the latent representations of reasoning texts with their corresponding relational semantics. The training set consists of about 30,000 images, including approximately 46,000 base-set triplet samples, following the original data distribution. To alleviate predicate long-tail bias, we perform image-level soft-balanced sampling by assigning each image a weight based on the inverse frequencies of its ground-truth predicates, so that images containing more rare predicates are sampled more frequently, and resample with replacement at a fixed dataset size.

Model and Optimization. The embedding encoder is initialized from the pretrained Qwen3-VL-Embedding-2B model [20] (official release) and fine-tuned using LoRA, where the backbone parameters are frozen and low-rank adapters are applied to attention projection layers. We set the LoRA rank to 8, the scaling factor to 16, and the dropout rate to 0.1. Text representations are obtained by selecting the last valid token (according to the attention mask), followed by ℓ_2 normalization along the embedding dimension. We train the model using AdamW with a learning rate of 2×10^{-4}, a cosine learning rate schedule, and warmup. Unless otherwise specified, we use a per-device batch size of 16, gradient accumulation of 2, and train for 3 epochs. Gradient clipping is applied with a maximum norm of 1.0.

Contrastive Learning. We optimize the embedding model using an InfoNCE-style contrastive objective with temperature $\tau = 0.1$. For a mini-batch of size N, each aligned pair (r_i, s_i) forms a positive pair, while all other labels in the batch serve as in-batch negatives. Similarity is computed via temperature-scaled dot product.

Connection to GRPO. After training, the embedding model is frozen and used as the encoder in the reward model for GRPO. Given a generated CoT, the encoder maps it into the embedding space, where its alignment with relational semantics is evaluated to compute the proposed representation-aligned specificity reward. This enables GRPO to optimize CoT generation toward semantically discriminative and relation-specific outputs.

C.3 GRPO for CoT

For GRPO training, we first randomly sample 2000 object-pair instances and use Qwen3-VL-30B-A3B-Instruct[1] to generate responses for supervised fine-tuning as a cold-start stage. We then sample approximately 20,000 object-pair instances for policy optimization. During GRPO, each object-pair prompt is decoded into a group of 4 responses with sampling temperature 0.7, and the group-relative advantage is computed from the proposed specificity reward. We implement GRPO based on the EasyR1 framework. The embedding reward model is kept frozen throughout policy optimization, and no novel predicate annotations are used during this stage. We train for one epoch with a learning rate of 10^{-6}, weight decay of 10^{-2}, no learning-rate warmup, and a gradient clipping norm of 1.0. We use PPO clipping bounds of $[0.8, 1.3]$, dual clipping with coefficient 3.0, and a low-variance KL penalty with weight 0.01. The specificity-reward temperature is 0.1, distinct from the rollout temperature of 0.7.

The format reward has weight 0.3 and equals one only when the response contains all three complete tag pairs: `<VISUAL_EVIDENCE>...</VISUAL_EVIDENCE>`, `<REASONING_THINKING>...</REASONING_THINKING>`, and `<FINAL_SUMMARY>...</FINAL_SUMMARY>`. Matching is case-insensitive and allows multiline content. Otherwise, both the format and specificity rewards are set to zero.

C.4 Additional Supervised Training on Embedding

To mitigate potential degradation in base-class performance after self-supervised training (e.g., reduced base mean recall), an additional supervised refinement stage is introduced.

Usage and Scope. The embedding model obtained from the self-supervised stage is frozen and used as the reward encoder in the GRPO training. In contrast, the supervised refinement stage is applied only to the embedding model used for downstream matching, and does not affect the GRPO optimization process. Specifically, the refined embedding is used at inference time to compute similarity between encoded CoT and candidate predicates.

Refinement Strategy. We perform supervised fine-tuning on the self-supervised embedding model using CoT-derived supervision. The training data is also restricted to base-set samples. Specifically, the full CoT texts are used as queries, while the corresponding ground-truth predicates (possibly multiple per sample) serve as labels. The construction of the training data follows the same procedure as in Sec. C.2, where image-level soft-balanced sampling is applied, and the dataset size is kept consistent.

Multi-positive Contrastive Learning. We adopt a multi-positive InfoNCE objective for training. Given a mini-batch, we first aggregate all predicate labels across samples and remove duplicates to form a candidate predicate set. Each query is associated with multiple positive predicates within this set, while the remaining predicates are treated as negatives. To improve computational efficiency, each unique predicate is encoded only once per batch. To prevent samples with a larger number of positives from dominating the training objective, we normalize the contributions of positive pairs by the number of positives for each query.

Model and Training Details. The embedding encoder is initialized from the self-supervised model and further fine-tuned using LoRA, where the backbone parameters remain frozen while the LoRA adapters are initialized from the self-supervised stage and remain trainable. The LoRA configuration is kept consistent with the previous stage (rank $r = 8$, scaling factor $\alpha = 16$, and dropout $= 0.1$). Text representations are obtained using mean pooling. Training is performed with AdamW using a learning rate of 1×10^{-4} and a cosine learning rate schedule with warmup. Unless otherwise specified, we use a per-device batch size of 16 with gradient accumulation of 2, and train for 3 epochs. The temperature parameter is set to $\tau = 0.12$. We use the final checkpoint for the GRPO-optimized MLLM and select embedding checkpoints with the lowest loss on a validation subset held out from the training set.

SGDet Inference and Evaluation. We use the candidate object pairs selected by the upstream detector. Under graph-constraint evaluation, we retain only the highest-scoring predicate for each directed object pair and rank triplets solely by relation scores, without incorporating object detection confidence.

Predicate Templates. We use Gemini 3 to generate approximately 3–7 textual templates for each predicate, which are used to construct the multi-template predicate prototypes.

C.5 Score Normalization and Prior Calibration

During inference, we first compute the predicate retrieval score s_k following Sec. 3.5. To reduce scale variation across different subject-object pairs, we apply pair-level z-score normalization over all predicate scores:

$$\hat{s}_k = \frac{s_k - \mu}{\max(\sigma, \epsilon)}, \tag{11}$$

where μ and σ are computed over all predicates for the current subject-object pair.

We then apply a lightweight subject-object prior calibration. For base predicates, the prior is estimated from training-set co-occurrence statistics with Dirichlet smoothing. For novel predicates, since no training annotations are available, we transfer the base prior according to predicate semantic similarity, where each novel predicate uses the mean prior of its top-K nearest base predicates in the predicate embedding space. We set $K = 5$ in all experiments.

The final calibrated score is:

$$\tilde{s}_k = \hat{s}_k + \eta \log \big(q(p_k \mid c_i, c_j)\big), \tag{12}$$

where q denotes the base prior for base predicates and the transferred prior for novel predicates. We set $\eta = 0.05$ in all experiments, and predict the relation by $\arg\max_k \tilde{s}_k$. This calibration is used only during inference.

D Supplementary Experiment: Zero-Shot PredCls Performance of Different MLLM Backbones

D.1 Experimental Setup

This experiment is conducted to verify the impact of the basic vision-semantic alignment capability of the MLLM backbone on the zero-shot performance of open-vocabulary relation reasoning.

The four off-the-shelf backbones are evaluated without fine-tuning, using the same prompt, decoding strategy, and TP+PR retrieval pipeline. PR uses base-training statistics. Trained RaRe is included as a reference.

D.2 Experimental Results

We adopt the standard OV-SGG evaluation metrics, and report the global recall R@50/100 and category-averaged mean recall mR@50/100 on the full set (Total), base classes (Base), and novel classes (Novel). The results are summarized in Table 6.

Table 6: PredCls results of four off-the-shelf MLLM backbones using TP+PR without parameter updates, with trained RaRe included as a reference. PR uses base-training statistics.

MLLM Backbone	**Total**		**Base**		**Novel**	
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
Qwen2.5-VL-3B-Instruct	48.86/52.11	17.35/19.96	49.29/52.57	19.68/22.35	9.64/11.29	11.92/14.39
Qwen3-VL-4B-Instruct	52.90/55.68	20.68/23.54	53.40/56.21	23.37/25.93	12.71/13.67	14.39/17.95
Qwen3-VL-8B-Instruct	53.13/55.96	21.60/23.94	53.57/56.42	23.63/25.89	15.56/17.34	16.87/19.41
Qwen3.5-4B	53.22/56.08	21.72/23.96	53.69/56.58	23.75/26.30	13.22/14.34	16.98/18.50
RaRe	49.90/51.33	24.62/25.80	50.36/51.81	25.53/26.45	12.75/13.60	22.49/24.27

The experimental results show that the MLLM backbone with stronger basic vision-semantic understanding capability exhibits better comprehensive performance on the zero-shot open-vocabulary relation reasoning task, especially stronger generalization ability on unseen novel categories. This result provides a basic reference for the backbone selection and optimization of subsequent MLLM-based OV-SGG methods.

E Sensitivity analysis

All experiments in this section are conducted on the same trained RaRe checkpoint under the PredCls setting. We use a fixed configuration of $K = 5$ and $\eta = 0.05$, chosen before test-set evaluation.

E.1 Sensitivity to the Number of Transferred Predicates

The sensitivity to the number of transferred base predicates is investigated by varying $K \in \{1, 3, 5, 10\}$, with the results summarized in Table 7. The optimal Total and Novel mR are attained at $K = 5$, whereas performance remains largely stable for $K \in \{3, 5, 10\}$. The degradation observed at $K = 1$ suggests that the prior transferred from a single neighbor is more susceptible to semantic mismatch, and is therefore less reliable for calibration.

E.2 Sensitivity to the Prior Weight

The sensitivity to the subject–object prior weight η in Eq. (12) is further examined, with the results reported in Table 8. The results indicate that the prior should be retained as a low-weight calibration signal: a modest weight yields a more favorable overall balance, whereas an excessive weight ($\eta = 0.1$) leads to a substantial degradation in Novel mR. This observation substantiates the treatment of the prior as a lightweight calibration term, such that relation-semantic scores remain the dominant

Table 7: Sensitivity analysis of the number of transferred base predicates K. All results are reported under the PredCls setting with $\eta = 0.05$ and uniform averaging transfer.

K	Total mR@50/100	Base mR@50/100	Novel mR@50/100
1	22.86/24.53	24.94/26.60	18.02/19.72
3	23.61/25.31	24.80/26.48	20.82/22.58
5	**24.62/25.80**	**25.53**/26.45	**22.49/24.27**
10	23.95/25.48	24.94/**26.69**	21.65/22.65

prediction signal. Among the tested values, $\eta = 0.01$ achieves the highest Novel mR, whereas the predefined setting $\eta = 0.05$ achieves the highest Total and Base mR.

Table 8: Sensitivity analysis of the subject–object prior weight η. All results are reported under the PredCls setting with $K = 5$ and uniform averaging transfer.

η	Total mR@50/100	Base mR@50/100	Novel mR@50/100
0	23.03/24.33	22.12/23.32	25.15/26.71
0.01	24.50/25.65	23.54/24.85	**26.75/27.50**
0.05	**24.62/25.80**	**25.53/26.45**	22.49/24.27
0.1	19.39/21.06	23.03/24.79	10.92/12.37

E.3 Comparison of Prior Transfer Strategies

Table 9 compares uniform averaging with distance-weighted transfer and with the removal of novel-prior transfer. Without transfer, base priors are retained and novel priors are fixed to $0.05/15$. This penalizes novel scores, reducing Novel mR@50 to 2.33 and highlighting the sensitivity to base–novel calibration.

Table 9: Comparison of novel-predicate prior transfer strategies. All results are reported under the PredCls setting with $K = 5$ and $\eta = 0.05$.

Method	Total mR@50/100	Base mR@50/100	Novel mR@50/100
Uniform	**24.62/25.80**	25.53/26.45	**22.49/24.27**
Distance-weighted	23.60/25.30	24.82/26.56	20.73/22.35
No novel-prior transfer	18.99/20.13	**26.14/27.56**	2.33/2.78

E.4 Examples of Transferred Neighbors

Table 10 presents examples of novel-prior transfer under the VG150 PredCls split, with neighbors retrieved exclusively from its base predicate vocabulary

F Comparison with Alternative MLLM Optimization Strategies

To compare optimization strategies under an identical backbone, an SFT-only checkpoint and a Relation-R1-style GRPO checkpoint were trained with same Qwen3-VL-4B-Instruct and evaluated under the same protocol. The SFT baseline is supervised with ground-truth base predicates. The GRPO baseline adopts a simplified form of the Relation-R1 reward [18]: since the PredCls setting provides object labels and boxes and queries a single object pair per sample, the original triplet-level criterion reduces to predicate correctness,

$$r = r_{\text{format}} + r_{\text{pred}},$$

where $r_{\text{pred}} = 1$ if the predicted predicate matches the ground-truth predicate and 0 otherwise.

Since only base-predicate annotations are available under the standard OV-SGG protocol, both SFT and correctness-based GRPO directly optimize a finite base-label space. This improves base-category recognition but biases the model toward base predicates and weakens novel-predicate generalization (Table 11).

Table 10: Examples of the top-5 nearest base predicates used for novel-prior transfer under the VG150 PredCls split.

Novel predicate	Top-5 nearest base predicates
across	1) to, 2) over, 3) in front of, 4) against, 5) behind
laying on	1) on, 2) sitting on, 3) under, 4) covered in, 5) attached to
on back of	1) on, 2) behind, 3) against, 4) over, 5) attached to

RaRe instead avoids predicate correctness during policy optimization and optimizes object-pair specificity in the relational representation space, which prevents direct attraction toward the base-label space. The stronger and more balanced mR of RaRe indicates additional benefits over conventional SFT and correctness-based GRPO.

Table 11: Same-backbone comparison of MLLM optimization strategies under the PredCls setting on the full VG150 test set. GRPO uses a binary predicate-correctness reward adapted from Relation-R1, while RaRe uses the specificity reward in Eq. (8).

	Total (Relation)		Base (Relation)		Novel (Relation)	
Model	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
SFT	**55.92/58.49**	16.83/18.70	**56.53/59.13**	20.69/22.80	7.33/8.13	7.84/9.14
GRPO	55.05/57.64	20.18/22.52	55.60/58.22	23.56/26.03	10.34/10.97	12.30/14.33
RaRe (Ours)	49.90/51.33	**24.62/25.80**	50.36/51.81	**25.53/26.45**	**12.75/13.60**	**22.49/24.27**

G Supplementary Experiment: SGCls Setting Results

In this section, we provide the complete experimental results of our method under the SGCls (Scene Graph Classification) setting, as supplementary validation of our method's robustness.

Table 12: **Experimental results of OVR-SGG on SGCls setting.**

Model	**Total (Relation)**		**Base (Relation)**		**Novel (Relation)**	
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
SGTR+	21.42/26.60	9.13/11.58	32.18/36.11	13.12/16.88	0.00/0.00	0.00/0.00
SGTR+ + RAHP	26.65/30.83	12.98/**16.81**	32.63/36.83	13.86/**17.14**	14.37/**21.72**	9.91/15.04
ACC+Ours	**30.46/33.56**	**13.76**/16.62	**33.79/37.39**	**14.30**/17.02	**15.50**/17.38	**12.51/15.67**

Under the SGCls setting, our method achieves 12.51% on Novel mR@50, outperforming the strongest baseline SGTR+ + RAHP [23] (9.91%).

H Cross-Task Transfer: Training-Free Zero-Shot HOI

We examine whether our relation reasoning pipeline can be adapted to zero-shot HOI using an off-the-shelf MLLM without task-specific training or parameter updates.

H.1 Experimental Setup

We use the off-the-shelf Qwen3-VL-4B-Instruct and Qwen3-VL-Embedding-2B models without SGG- or HOI-specific fine-tuning. Human–object pair construction and object-conditioned HOI candidate filtering follow the training-free configuration of ADA-CM [17], using its DETR-R50 detector. We apply our CoT-based embedding-retrieval pipeline to score the candidate interactions.

Experiments are conducted on HICO-DET [2], the standard benchmark for ZS-HOI. We follow the general ZS-HOI protocol and use mean Average Precision (mAP) as the evaluation metric, reporting results on the full set (Full), rare categories (Rare), and non-rare categories (Non-rare). We compare our method with state-of-the-art ZS-HOI methods under the exact same training-free setting, including UPT (TF) [40], ADA-CM [17] (ICCV 2023), and DA-HOI [35] (ICLR 2026).

H.2 Results and Conclusion

The experimental results are summarized in Table 13.

Table 13: Training-Free Zero-Shot HOI results on HICO-DET (metric: mAP). All methods are under the fully training-free setting.

Method	Full	Rare	Non-rare
UPT (TF)	13.64	9.66	14.85
ADA-CM (ICCV 2023)	25.19	27.24	24.58
DA-HOI (ICLR 2026)	**31.50**	-	-
Ours	27.71	**29.06**	**27.31**

Our pipeline achieves 27.71 Full mAP and 29.06 Rare mAP on HICO-DET without task-specific training.These results provide evidence of cross-task transfer to zero-shot HOI.

I Additional Analysis on Representation Properties

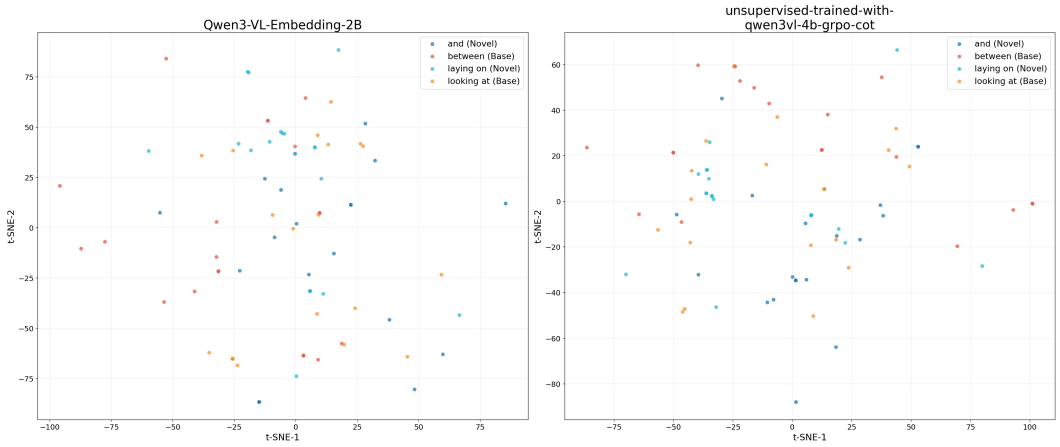


Figure 5: t-SNE visualization of CoT representations projected by Baseline (Qwen3-VL-Embedding-2B), and +CoT+GRPO.

Table 14: **Discriminability of Unsupervised-trained Embedding** We report Silhouette Score and macro_mean_ratio for Global, Base, and Novel subsets. ↑ Δ denotes improvement over the baseline. Baseline uses the off-the-shelf embedding model and CoTs from the unoptimized MLLM. +CoT replaces only the embedding model with its SSL-trained counterpart. +CoT+GRPO retains the same SSL-trained embedding model and uses CoTs from the GRPO-optimized MLLM.

Metric	Baseline	+CoT	+CoT+GRPO
Silhouette (Global)	-0.1313	-0.0935 (↑0.0378)	-0.0924 (↑0.0389)
Silhouette (Base)	-0.1040	-0.0689 (↑0.0351)	-0.0657 (↑0.0383)
Silhouette (Novel)	+0.0298	+0.0514 (↑0.0216)	+0.0511 (↑0.0213)
macro_mean_ratio (Global)	0.0944	0.2876 (↑0.1932)	0.3177 (↑0.2233)
macro_mean_ratio (Base)	0.0863	0.2766 (↑0.1903)	0.3071 (↑0.2208)
macro_mean_ratio (Novel)	0.9683	1.4294 (↑0.4611)	1.4717 (↑0.5034)

Discriminability Analysis Since qualitative clustering visualizations in Figure5 are insufficient to capture fine-grained inter-predicate variation, we adopt two quantitative metrics: Silhouette Score, which measures intra-class compactness relative to inter-class separation (range $[-1, 1]$, higher is better), and **macro _mean _ratio**, the weighted mean of nearest inter-class distance over intra-class

scatter per category (values > 1 indicate basic discriminability). Both metrics are calculated on a dedicated evaluation subset of 2,000 images containing 12,686 relational triplets, spanning both base and novel predicates. We utilize CoT representations generated by Qwen3-VL-4B-Instruct across different embedding models to assess how our training strategy enhances the relational separability in the embedding space, as summarized in Table 14.

Both trained variants improve over the baseline on all reported metrics. Compared with CoT-only training, the GRPO variant further improves most metrics, while Novel Silhouette remains nearly unchanged (0.0514 versus 0.0511). Global and Base Silhouette scores remain negative, indicating improved but still limited cluster separation. Overall, these results suggest improved relational discriminability, including for novel predicates.

The metrics are defined as follows:

(1) *Silhouette Score* for a sample i is given by:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{13}$$

where $a(i) = \frac{1}{|C_I|-1} \sum_{j \in C_I, i \neq j} d(i,j)$ is the average intra-class distance, and $b(i) = \min_{J \neq I} (\frac{1}{|C_J|} \sum_{j \in C_J} d(i,j))$ is the average distance to the nearest neighboring cluster.

(2) **macro_mean_ratio** for a set of predicates is defined as:

$$\text{macro_mean_ratio} = \frac{\sum_{k=1}^{K} n_k \cdot R_k}{\sum_{k=1}^{K} n_k}, \quad R_k = \frac{d_{cos}(\mu_k, \mu_{nearest})}{\sigma_k + \tau} \tag{14}$$

where n_k is the number of samples for predicate C_k, and R_k is the stabilized discriminability ratio. $d_{cos}(\mu_k, \mu_{nearest})$ denotes the cosine distance between the centroid of C_k and its nearest neighbor centroid, while σ_k is the mean cosine distance of samples in C_k to their own centroid. To prevent divergence when $\sigma_k \to 0$, we introduce a stabilization term $\tau = \max(10^{-6}, 0.1 \times \text{median}(\{\sigma\}_K))$. Despite its name, macro_mean_ratio is weighted by category frequency rather than averaged equally across categories.

CoT Length Table 15 compares the average CoT length on the full VG150 test set under PredCls. RaRe reduces the average length from 392.8 to 297.1 tokens, suggesting that its gains do not rely on longer generated rationales.

Table 15: **Average CoT length on the full VG150 test set under PredCls.**

Model	Average Length (tokens)
Qwen3-VL-4B-Instruct	392.8
RaRe	297.1

J Additional Controls and Inference Cost

Object-only control. To verify that the performance relies on visual input rather than on object annotations alone, the checkpoint is re-evaluated with the image removed, while object labels, boxes, prompt, and inference pipeline are kept unchanged (Table 17). Performance drops consistently across all three groups, and the drop is largest on novel predicates (from 22.49 to 13.71 at mR@50), indicating that object annotations alone do not account for the reported results.

Direct output vs. embedding retrieval. Using the same off-the-shelf Qwen3-VL-4B-Instruct model and prompt, direct predicate prediction is compared with the training-free embedding-retrieval pipeline (Table 18). Retrieval improves Total, Base, and Novel mR@100 by 8.05, 5.77, and 13.37 points, respectively. Since the MLLM and the prompt are identical in both variants, the gap is attributable to the embedding-based retrieval, which also serves as the reward encoder in Stage 2 rather than being introduced only at inference time.

Table 16: **Fixed-proposal SGDet control on the full VG150 test set.** Both configurations use identical ACC object proposals and candidate pairs. The RaRe configuration adds RL, SSL, and PR to training-free retrieval; ST is not used.

Configuration	Total mR@50/100	Novel mR@50/100
Training-free retrieval	10.08/12.19	10.21/12.27
+ RL + SSL + PR	10.38/12.32	6.69/8.02

Fixed-proposal SGDet control. Table 16 compares two configurations using identical ACC boxes, object labels, candidate pairs, decoding settings, candidate filtering, and TP-based retrieval. Adding RL, SSL, and PR yields modest Total mR@50/100 gains of 0.30/0.13 points, while reducing Novel mR@50/100 by 3.52/4.25 points. Together with the PSG ablation in Table 5, this indicates that the benefits of learning and calibration vary across evaluation settings.

Inference cost. Table 19 reports the per-image latency of the SGDet pipeline on four NVIDIA L40 GPUs. CoT generation dominates the cost (3.526s, 64.7%), followed by object detection (1.247s, 22.9%), while embedding retrieval accounts for 0.677s.

Table 17: Object-only control under the PredCls setting on the full VG150 test set. Only the image is removed; object labels, boxes, prompt, and inference pipeline are unchanged.

Variant	Total mR@50/100	Base mR@50/100	Novel mR@50/100
RaRe w/o image	16.78/18.02	18.10/19.50	13.71/14.56
RaRe	**24.62/25.80**	**25.53/26.45**	**22.49/24.27**

Table 18: Training-free comparison between direct MLLM output and embedding retrieval under the PredCls setting on the full VG150 test set. Both variants use the same MLLM and prompt.

Method	Total mR@100	Base mR@100	Novel mR@100
Direct MLLM output	12.20	13.25	9.76
Embedding retrieval	20.25	19.02	23.13
Improvement	+8.05	+5.77	+13.37

K Long-Tailed Recognition and Novel-Predicate Coverage under SGDet

Base-Predicate Long-Tailed Recognition. To distinguish training-frequency effects from novel-predicate generalization, we rank the 35 base predicates by their frequency in the VG150 training set and divide them into head, body, and tail groups using a 20%/40%/40% class split. Both methods are evaluated on the full VG150 test set with identical ACC proposals under graph-constraint evaluation. Table 20 shows that RaRe substantially improves body and tail recognition, with a small decrease on head predicates. Under the same top-100 prediction budget, it also allocates larger shares of base predictions to body and tail predicates (Table 21). The accompanying recall gains indicate that this broader coverage translates into improved recognition of less frequent predicates.

Novel-Predicate Coverage. All statistics are computed on the full VG test set with the fixed ACC proposals of the SGDet evaluation in Table 1, where Ours denotes the RaRe variant. The 15 OvR novel predicates are ranked by ground-truth frequency and split 20%/40%/40% into three head predicates (on, of, and in), six body predicates, and six tail predicates; the head group holds 92.4% of the novel GT instances, the body group 6.9%, and the tail group 0.7%. Since mR averages over predicates, each of the 12 body and tail predicates weighs as much as on.

Table 22 reports the prediction counts per group: 942k, 0, and 124 for ACC and 644k, 77k, and 70k for RaRe, with an identical total of 2,644,600 predictions and a smaller novel share for RaRe (29.93% vs. 35.64%). At mR@50, ACC obtains 0.124 on the head group and 0.000 on the other two, whereas RaRe obtains 0.083, 0.062, and 0.064. Per predicate (Table 23), RaRe is lower on on (0.103

Table 19: Inference-time breakdown of the SGDet pipeline on four NVIDIA L40 GPUs, reported per image.

Component	Time (s/image)	Share
Object detection (ACC)	1.247	22.9%
CoT reasoning generation	3.526	64.7%
Embedding retrieval	0.677	12.4%
Total	5.450	100.0%

Table 20: **Frequency-stratified results on VG150 base predicates under SGDet.** Head, body, and tail contain 7, 14, and 14 predicates, respectively, ranked by training frequency.

Method	Head mR@50/100	Body mR@50/100	Tail mR@50/100
ACC	**24.0/28.0**	0.7/0.9	0.0/0.0
RaRe	22.8/26.7	**14.3/16.3**	**4.2/5.7**

vs. 0.273) and higher on of (0.078 vs. 0.072), in (0.068 vs. 0.027), and 7 of the 12 body and tail predicates, including riding (0.310), eating (0.134), and says (0.250).

Five predicates stay at zero for both models at R@50 (wears, part of, walking in, painted on, and playing; part of reaches 0.004 for RaRe at R@100), and some of them receive far more predictions than they have GT instances (says: 12 GT instances, 44k predictions), so their recall rests on few positives. The gain thus comes from categories that ACC leaves at zero, each worth 1/15 of the mR, whereas the loss on on is bounded by a single 1/15; the remaining gap is set by the five predicates that neither model recovers.

Table 21: **Prediction distribution across base-predicate frequency groups.** Base pred. denotes the share of all top-100 predictions assigned to base predicates. Head, body, and tail shares are normalized within base predictions.

Method	Base pred. (%)	Head (%)	Body (%)	Tail (%)
ACC	64.36	99.4	0.6	≈ 0.0
RaRe	70.07	72.9	19.2	7.9

Table 22: Novel-predicate recall and prediction counts by head/body/tail group under SGDet with fixed ACC proposals. All results are reported at mR@50/100 on the full VG test set.

Bucket	GT share	Ours@50	ACC@50	Ours@100	ACC@100	Pred. (Ours/ACC)
Head	92.4%	0.083	0.124	0.112	0.159	644k/942k
Body	6.9%	0.062	0.000	0.074	0.000	77k/0
Tail	0.7%	0.064	0.000	0.070	0.000	70k/124

Table 23: Per-predicate novel recall, test-set frequency, and prediction counts under the VG150 SGDet split with fixed ACC proposals. Pred. # reports the number of predictions, and paired entries correspond to Ours/ACC

			R@50		R@100		Prediction (Ours/ACC)	
Predicate	Bucket	Test#	Ours	ACC	Ours	ACC	#	%
on	Head	50947	0.103	0.273	0.119	0.334	409k/827k	15.47/31.26
of	Head	17074	0.078	0.072	0.129	0.110	81k/95k	3.06/3.58
in	Head	9138	0.068	0.027	0.089	0.031	154k/21k	5.82/0.79
wears	Body	2059	0.000	0.000	0.000	0.000	1.7k/0	0.06/0.00
riding	Body	1348	0.310	0.000	0.352	0.000	26k/0	0.99/0.00
attached to	Body	742	0.036	0.000	0.058	0.000	21k/0	0.78/0.00
belonging to	Body	666	0.007	0.000	0.015	0.000	8.6k/0	0.33/0.00
walking on	Body	577	0.006	0.000	0.009	0.000	11k/0	0.40/0.00
for	Body	335	0.010	0.000	0.010	0.000	9.2k/0	0.35/0.00
eating	Tail	232	0.134	0.000	0.168	0.000	6.6k/0	0.25/0.00
part of	Tail	138	0.000	0.000	0.004	0.000	8.1k/0	0.31/0.00
walking in	Tail	105	0.000	0.000	0.000	0.000	5.2k/1	0.20/0.00
painted on	Tail	96	0.000	0.000	0.000	0.000	2.2k/0	0.08/0.00
playing	Tail	28	0.000	0.000	0.000	0.000	3.8k/0	0.15/0.00
says	Tail	12	0.250	0.000	0.250	0.000	44k/123	1.68/0.00

L More Qualitative Comparison Result

Predictions are obtained by embedding retrieval for the queried object pair and may differ from the generated summary. Correctness is determined by matching the ground-truth predicate; a correct prediction does not imply that every statement in the rationale is visually accurate.



Qwen3-vl-4b-instruct (Untrained):
CoT: <VISUAL_EVIDENCE> The image shows a row of motorcycles parked on a paved street in front of a building with "PARTS" and "SALES" signs. The motorcycles are aligned parallel to the street, positioned on the asphalt surface. The street occupies the lower portion of the image, extending from the left edge to the right, with the motorcycles resting directly on it. The building is behind the motorcycles, and the street is clearly the ground surface upon which the motorcycles are situated.
< /VISUAL_EVIDENCE>
<REASONING_THINKING> Step 1: Identify the objects. The two objects are a motorcycle and a street. The motorcycle is visible in the foreground, parked on the street. Step 2: Analyze spatial relationship. The motorcycle is positioned on the surface of the street, which is a paved road. The street is the ground plane, and the motorcycle is resting on it. Step 3: Determine interaction mechanism. The motorcycle is not floating or suspended; it is placed on the street surface, indicating a physical resting relationship. The street provides a stable, flat surface for the motorcycle to be parked on. Step 4: Confirm no other relationships. There is no indication of the motorcycle being attached to the building, or the street being above the motorcycle. The motorcycle is clearly on the street, not beside or behind it. Step 5: Conclude relationship. The motorcycle is resting on the street, which is the most direct and logical spatial relationship based on the visual evidence. < /REASONING_THINKING>
<FINAL_SUMMARY> motorcycle on street The motorcycle is parked on the street surface.
< /FINAL_SUMMARY>
Prediction: **motorcycle on street** ×

Ours :
CoT: <VISUAL_EVIDENCE> The image shows a motorcycle positioned on a paved surface, which appears to be a street. The motorcycle is stationary, with its wheels in contact with the ground, and it is located near the edge of the street. The street is a long, flat expanse extending horizontally across the frame, and the motorcycle is clearly placed on it rather than on a sidewalk or other surface. The surrounding area suggests an outdoor urban or suburban setting. < /VISUAL_EVIDENCE>
<REASONING_THINKING> The motorcycle is resting on the surface of the street, with no visible signs of motion or active use. Its wheels are aligned with the street's edge, and there is no indication of it being in motion. The position and orientation of the motorcycle suggest it is not currently being driven. The lack of movement and the environment imply that the motorcycle is not operating at the moment. The street provides a flat, stable surface suitable for the motorcycle to be parked on. The interaction between the motorcycle and the street is consistent with a scenario where the motorcycle is temporarily placed in a stationary position. < /REASONING_THINKING>
<FINAL_SUMMARY> motorcycle parked on street The motorcycle is stationary on the surface of the street, indicating it is not in motion and is instead resting on the street. < /FINAL_SUMMARY>
Prediction: **motorcycle parked on street** ✓



Qwen3-vl-4b-instruct (Untrained):
CoT: <VISUAL_EVIDENCE> The image shows two individuals in wetsuits standing in shallow ocean water, with waves breaking behind them. The man on the left is holding a surfboard under his arm, which is predominantly white with green and black markings. The surfboard is positioned diagonally across his body, extending from his left shoulder to his right hip. The second individual, to the right, is also holding a surfboard, which is partially obscured by his body and appears to be tucked under his left arm. The surfboard he holds is mostly white with a visible black leash extending toward the right side of the frame. Both individuals are facing forward, with their bodies angled slightly toward the camera, and are partially submerged in water, creating splashes around their legs. < /VISUAL_EVIDENCE>
<REASONING_THINKING> Step 1: Identify the two objects and their respective holders. The man on the left is holding a surfboard, and the second individual is also holding a surfboard. Both are in wetsuits, suggesting they are surfers. Step 2: Analyze the spatial relationship. The surfboard held by the man on the left is clearly visible and positioned under his arm, with its length extending diagonally across his torso. The surfboard held by the second individual is similarly held under the arm, though partially obscured. Both surfboards are being carried, not actively used for surfing at this moment. Step 3: Determine the interaction mechanism. The interaction is one of carrying or transporting the surfboards. The surfboards are not in use for riding waves but are being carried by the individuals, likely preparing to enter the water or having just exited. Step 4: Consider context. The presence of ocean waves and the individuals' wetsuits reinforce the context of surfing. The fact that both are holding surfboards and standing in shallow water suggests they are either preparing to surf or have just finished. Step 5: Conclude the relationship. The predicate relationship is that the man is carrying a surfboard, and the second individual is also carrying a surfboard. The relationship is mutual and symmetric in nature, as both are engaged in the same action of carrying surfboards. < /REASONING_THINKING>
<FINAL_SUMMARY> man carries surfboard, second individual carries surfboard. Both individuals are carrying surfboards while standing in shallow ocean water, likely preparing for or having just finished surfing. < /FINAL_SUMMARY>

Prediction: man on surfboard ×

Ours :
CoT: <VISUAL_EVIDENCE> The image shows a man standing near a surfboard. The surfboard is positioned horizontally and appears to be held or carried by the man, with one end near his body and the other extending outward. The man is facing forward, and the surfboard is aligned parallel to his torso. The surfboard's orientation suggests it is being carried rather than being placed on the ground or used for surfing. < /VISUAL_EVIDENCE>
<REASONING_THINKING> The man is holding the surfboard in a way that indicates he is transporting it. The surfboard is not floating or being ridden, nor is it placed on the ground. Its position relative to the man—aligned with his body and held close—suggests it is being carried. This indicates an interaction where the man is moving the surfboard with him, likely for transport to or from a beach. The alignment and grip imply a structured method of carrying, consistent with a person holding an object.
< /REASONING_THINKING>
<FINAL_SUMMARY> man carrying surfboard The man is transporting the surfboard by holding it horizontally along his body. < /FINAL_SUMMARY>

Prediction: man carrying surfboard ✓



Qwen3-vl-4b-instruct (Untrained):
CoT: <VISUAL_EVIDENCE> The skateboard is positioned beneath the person's feet, with the person standing on it and actively riding. The wheel is one of the four wheels attached to the underside of the skateboard, visible in the lower portion of the skateboard's structure. The skateboard is oriented horizontally on the asphalt surface, and the wheel is part of this assembly, directly connected to the board's deck. The image is taken from behind the rider, with strong backlighting creating a bright flare, but the skateboard and wheel remain clearly identifiable in the foreground. < /VISUAL_EVIDENCE>
<REASONING_THINKING> Step 1: Identify the skateboard as the primary object being ridden. The person is standing on it, indicating it is the platform for movement. Step 2: Observe the wheel as a component of the skateboard. It is attached to the board's underside and is part of the structural assembly that allows the skateboard to roll. Step 3: Establish the relationship: the wheel is physically attached to the skateboard, and the skateboard is being used by the person. The wheel's function is to support the skateboard's movement on the ground, and it is inseparable from the skateboard's structure. Step 4: The interaction mechanism is mechanical: the wheel rolls on the pavement, enabling the skateboard to move. The skateboard contains the wheel as an integral part, and the person interacts with the skateboard via standing on it, which in turn causes the wheel to rotate and propel the skateboard forward. Step 5: Therefore, the wheel is a part of the skateboard, and the skateboard is being used by the person to ride. The relationship is not of direct interaction between the wheel and the person, but rather a structural and functional relationship between the wheel and the skateboard, with the person using the skateboard.
< /REASONING_THINKING>
<FINAL_SUMMARY> wheel is part of skateboard The wheel is an integral component of the skateboard, enabling its movement on the ground as the person rides it. < /FINAL_SUMMARY>

Prediction: **skateboard part of wheel** ×

Ours :
CoT: <VISUAL_EVIDENCE> The image shows a skateboard positioned horizontally, with a wheel attached to its underside. The wheel is located near the front of the skateboard and appears to be connected to the deck via an axle, consistent with the typical mechanical structure of a skateboard. The wheel is aligned with the front edge of the deck, suggesting a functional integration rather than a decorative placement. < /VISUAL_EVIDENCE>
<REASONING_THINKING> The skateboard is a two-wheeled vehicle designed for motion on flat surfaces. The wheel is visibly connected to the underside of the deck, near the front, and its position and attachment style match the standard configuration of a skateboard's front wheel assembly. This arrangement allows the wheel to rotate and engage with the ground, enabling forward movement. The spatial relationship indicates that the wheel is not merely placed nearby but is structurally integrated into the skateboard's motion system, consistent with the function of a front wheel on a skateboard.
< /REASONING_THINKING>
<FINAL_SUMMARY> The skateboard has a front wheel. The wheel is attached to the front of the skateboard and functions as part of its movement system. < /FINAL_SUMMARY>

Prediction: **skateboard has wheel** ✓

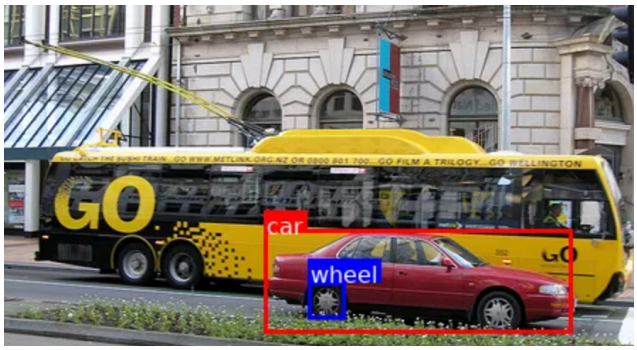
Bad case



Ours:
CoT: <VISUAL_EVIDENCE> The image shows a man positioned on a skateboard, with his body leaning forward and his arms slightly extended. The man's feet are placed on the deck of the skateboard, and the skateboard is oriented horizontally beneath him. The man appears to be actively engaged in a motion, possibly balancing or moving forward, with his upper body slightly bent. The overall posture suggests dynamic movement, and the man is clearly in control of the skateboard's direction and speed.
< /VISUAL_EVIDENCE>
<REASONING_THINKING> The man is on the skateboard, which is a surface that allows for motion on the ground. The positioning of his feet and body indicates that he is using the skateboard as a means of transportation. His forward-leaning posture and extended arms suggest an effort to maintain balance while moving. The interaction between the man and the skateboard involves the man applying force to the deck with his feet, causing the skateboard to accelerate and navigate forward. This interaction is consistent with the man using the skateboard to propel himself forward and control the motion of the device.
< /REASONING_THINKING>
<FINAL_SUMMARY> The relationship predicate is: man riding skateboard. The man is using the skateboard as a vehicle to move forward across the surface. < /FINAL_SUMMARY>

Prediction:
hand of man ×
GT Prediction: on

Bad case



Ours:
CoT: <VISUAL_EVIDENCE> The image shows a car positioned horizontally, with a wheel located beneath it and slightly to the left. The wheel is smaller in size compared to the car and appears to be attached to the lower portion of the car's body. The wheel is aligned with the car's undercarriage, indicating a structural connection. The spatial arrangement suggests that the wheel is not independently placed but is instead integrated into the car's structure.< /VISUAL_EVIDENCE> <REASONING_THINKING> The wheel is situated directly under the car and is oriented in a way that implies it is meant to rotate and support the car's movement. The wheel is connected to the car's chassis, which is typical of a wheel being mounted to a vehicle. The alignment and positioning suggest a functional integration where the wheel serves as a component that enables motion and support. This configuration indicates that the wheel is not simply adjacent to the car but is instead attached to it in a way that allows for movement and stability. The interaction mechanism involves the wheel being fixed to the car's frame, enabling it to rotate and move the vehicle forward or backward. < /REASONING_THINKING> <FINAL_SUMMARY> car mounted on wheel The wheel is attached to the car's underside, enabling it to rotate and support the vehicle's movement. < /FINAL_SUMMARY>

Prediction:
car mounted on wheel ×
GT Prediction: car has wheel

Bad case



Ours:
CoT: <VISUAL_EVIDENCE> The image shows a horse with a rider positioned on its back. The rider is seated on the horse's body, with their legs wrapped around the horse's sides and their hands likely holding the reins. The horse is standing still, and the rider appears to be in a stable, upright posture, indicating control and balance. The positioning suggests a coordinated interaction between the rider and the horse.< /VISUAL_EVIDENCE> <REASONING_THINKING> The rider is located on the horse's back, with their body aligned in a manner consistent with being seated. The proximity and posture of the rider suggest they are using the horse for transportation or control. The horse's stable stance and the rider's upright position imply a functional relationship where the rider is guiding or operating the horse. This coordination is typical of a controlled interaction involving movement or direction, which aligns with the idea of the rider being in control of the horse's motion.< /REASONING_THINKING> <FINAL_SUMMARY> The relationship is: rider operates horse. The rider is seated on the horse and is actively controlling its movement. < /FINAL_SUMMARY>

Prediction:
face riding horse ×
GT Prediction: on

Failure case of our method. The generated CoT contains hallucinated visual evidence and describes an incorrect object pair, leading to a mismatched relation prediction.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We reported the claim in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We reported the limitation as the end of paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper is not about theory.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided details about the methodology and implementation in the main paper and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and model will be released.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We reported the experimental setup and details in the main paper and appendix

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to limited computational resources, we were unable to run the experiments multiple times to compute error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Its reported in the paper and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully reviewed the NeurIPS Code of Ethics and confirm that the research conducted and presented in this paper fully complies with all ethical guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 5 briefly discusses potential benefits and downstream risks.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the datasets, pretrained models, and third-party code used in this work and have checked their licenses and complied with the applicable terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper did not introduce a new asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The MLLM is the core component of our framework. We have detailed the specific MLLMs used, the prompting strategies, and the generation pipeline in the paper.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.