Relation-R1: Progressively Cognitive Chain-of-Thought Guided Reinforcement Learning for Unified Relation Comprehension

Lin Li[1,2*], **Wei Chen**[1*], **Jiahui Li**[3], **Kwang-Ting Cheng**[1,2], **Long Chen**[1†]

[1] The Hong Kong University of Science and Technology
[2] AI Chip Center for Emerging Smart Systems
[3] Zhejiang University
lllidy, wchendb, timcheng, longchen@ust.hk, jiahuil@zju.edu.cn

Abstract

Recent advances in multi-modal large language models (MLLMs) have significantly improved object-level grounding and region captioning. However, they remain limited in visual relation understanding, struggling even with binary relation detection, let alone *N*-ary relations involving multiple semantic roles. The core reason is the lack of modeling for *structural semantic dependencies* among multi-entities, leading to over-reliance on language priors (*e.g.*, defaulting to "person drinks a milk" if a person is merely holding it). To this end, we propose Relation-R1, the *first unified* relation comprehension framework that explicitly integrates cognitive chain-of-thought (CoT)-guided supervised fine-tuning (SFT) and group relative policy optimization (GRPO) within a reinforcement learning (RL) paradigm. Specifically, we first establish foundational reasoning capabilities via SFT, enforcing structured outputs with thinking processes. Then, GRPO is utilized to refine these outputs via multi-rewards optimization, prioritizing visual-semantic grounding over language-induced biases, thereby improving generalization capability. Furthermore, we investigate the impact of various CoT strategies within this framework, demonstrating that a specific-to-general progressive approach in CoT guidance further improves generalization, especially in capturing synonymous *N*-ary relations. Extensive experiments on widely-used PSG and SWiG datasets demonstrate that Relation-R1 achieves state-of-the-art performance in both binary and *N*-ary relation understanding.

Code — github.com/HKUST-LongGroup/Relation-R1

Introduction

Recent advancements in MLLMs have significantly enhanced holistic image understanding (Liu et al. 2023a; Zhu et al. 2024) and object-level grounding (You et al. 2023; Zhang et al. 2024; Rasheed et al. 2024; Yuan et al. 2024, 2025; Lai et al. 2024; Peng et al. 2024) capabilities, enabling tasks like region captioning and referring question answering. However, current MLLMs exhibit critical limitations in relational scene understanding – a core competency for achieving human-like visual cognition, as illustrated in Figure 1(a).

Despite notable progress in binary relation detection, *i.e.*, identifying subject-object interactions (*e.g.*, "`child-drinking-glass`"), current models (Wang et al. 2024; Xu et al. 2025) struggle to capture *N*-ary relationships. Accurately modeling such relationships requires models to identify multiple *N*-th entities engaged in the relation with their distinct semantic roles (*what*, *who*, and *where* concepts) (Pratt et al. 2020; Cho et al. 2021). The neglect of these crucial *structural semantic dependencies* among multi-entities, such as the glass's role as the functional `container` of milk in the drinking action, may result in suboptimal relation triplets, *e.g.*, "`child-drinking-glass`". Furthermore, these unreliable and superficial pair-wise relations may make the MLLMs prone to over-reliance on language priors rather than visual-semantic cues (*e.g.*, evidence from semantically grounded visual regions). Consequently, there arises an urgent demand for a unified framework capable of addressing these shortcomings. This raises a critical question: ***How can we devise a unified framework for joint binary and N-ary relational reasoning with grounded cues?***

Intuitively, such complicated relation comprehension requires both strong **reasoning** and **generalization** abilities. As for *reasoning*, this capability is essential as complex relations are often implicit and require the model to perform sophisticated inference by integrating diverse and often subtle visual cues. This involves a multi-step reasoning process, *e.g.*, entity identification (*e.g.*, recognizing "`glass`" category appears in the image), relation recognition (*e.g.*, recognizing "`drink`" action appears in the image), role inference (*e.g.*, "`glass`" as a container), localization, and contextual integration (*e.g.*, observing a child holding a glass near their mouth suggests the "`drink`" action is likely occurring). In terms of *generalization*, models trained on limited data struggle to extrapolate to novel interactions and entity combinations or ambiguous contexts (*e.g.*, "`cup`" as a container for "`drinking`" action). In addition, inductive bias from language priors (*i.e.*, common-sense assumptions learned purely from text that may not always align with the visual reality) often dominates over visual grounding, further limiting their robustness. For example, a model might struggle to recognize "`drinking`" from "`milk`" if its language training strongly associates this action with "`glass`".

Fortunately, DeepSeek-R1-Zero (DeepSeek-AI 2025) has successfully demonstrated the emergence of reasoning ca-

*These authors contributed equally.
†Long Chen is the corresponding author.

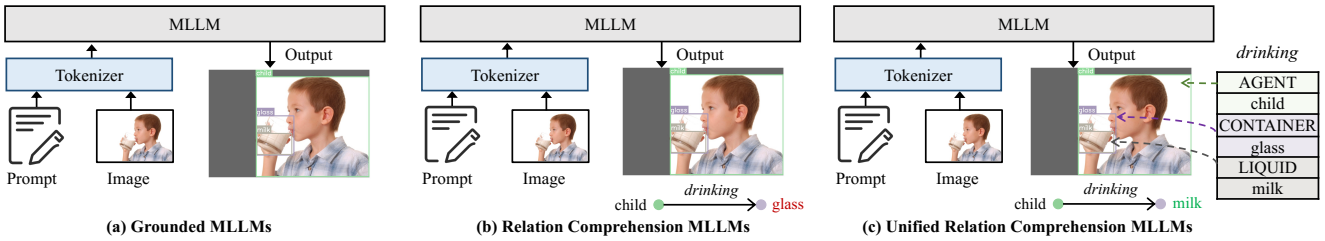

Figure 1: Illustrations of our Relation-R1 and other MLLMs. (a) **Grounded MLLMs**: Link objects in text to image regions but fail to capture object relations. (b) **Relation Comprehension MLLMs**: Model pairwise object relations but lack generalization to N-ary interactions. (c) **Unified Relation Comprehension MLLMs**: Jointly handle binary and N-ary relation detection.

pabilities in LLMs purely through RL. However, due to the inherent difficulty of reference models in understanding structural relationships, directly applying RL often struggles to produce consistently formatted outputs that are claimed in the input prompt (Chu et al. 2025). Conversely, when solely adopting supervised fine-tuning, the model suffers from poor generalization due to its overfitting to fixed training patterns and the limited diversity of annotated relational data.

To this end, inspired by the breakthrough of R1-series (DeepSeek-AI 2025; Yang et al. 2025; Zhao, Wei, and Bo 2025) in enabling reasoning and the ability of SFT to provide structured guidance, we propose a unified two-stage relation comprehension framework, **Relation-R1**, which combines the advantages of both **supervised fine-tuning** and **reinforcement learning** to empower MLLMs with relational reasoning and generalization capability:

Stage 1. SFT: Establishes foundational reasoning through step-by-step **cognitive chain-of-thought** guidance. This CoT enables the model to break down complex relation comprehension into a sequence of smaller, interpretable steps with visual evidence, such as object detection, spatial localization, *etc*. Thereby, the model learns to think based on grounded visual content over language inductive bias. Furthermore, its annotated relation-contained instructions ensure that the model generates outputs in a standardized format, facilitating the subsequent computation of rewards during RL.

Stage 2. RL: Refines the structured outputs generated by the SFT model through group relative policy optimization. The learning process is guided by multi-rewards: 1) format rewards, which incentivize the generation of thinking process (DeepSeek-AI 2025); 2) binary relation rewards, which encourage accurate identification of pairwise relationships between entities; and 3) N-ary relation rewards, which promote the comprehensive understanding of more complex, multi-entity activity. This stage employs policy gradient updates to cultivate the model to explore various potential solutions and optimizes its output based on defined multi-rewards, fostering robust relational reasoning and generalization.

Furthermore, to gain deeper insights into varying cognitive CoTs, we investigate the impact of template-based CoT and MLLM-generated CoT strategies on binary and N-ary relation comprehension, separately. Notably, by employing a progressive paradigm that initially adopts the template-based CoTs to guide the normative cognitive reasoning process, and then fine-tunes the model with the variable reasoning pathways introduced by a few MLLM-generated CoTs, our Relation-R1 demonstrates better generalization ability. Especially in the N-ary relations, the capacity to explore *synonymous relational expressions* has emerged.

We validate the effectiveness of Relation-R1 on both binary relation comprehension (*e.g.*, scene graph generation (Xu et al. 2017; Wang et al. 2024)) and N-ary relation comprehension task (*e.g.*, grounded situation recognition (Pratt et al. 2020)) across diverse experimental settings. Experiments on the PSG (Yang et al. 2022) and SWiG (Pratt et al. 2020) datasets demonstrate state-of-the-art performance in relational reasoning while maintaining generalization in ambiguous contexts. In summary, our contributions are fourfold:

- We reveal the limitations of the existing MLLMs in relation understanding, and incorporate both binary and N-ary relation detection into one unified relation comprehension framework.
- We propose Relation-R1, a unified framework that integrates cognitive CoT-guided SFT with RL. To the best of our knowledge, this is the first work that enables both relational reasoning and robust generalization in a deep-thinking fashion.
- We comparatively analyze template-based and MLLM-generated CoTs, demonstrating that progressively guiding the learning process with CoTs in a specific-to-general manner enhances generalization capabilities, particularly in capturing synonymous N-ary relations.
- Experimental results on the prevalent relation understanding datasets (*e.g.*, **+6.84**$\sim$**6.90**% gains on PSG dataset) under different settings demonstrate the effectiveness of Relation-R1.

Related Work

Reasoning in MLLMs. Recent LLMs have demonstrated advanced reasoning capabilities by emulating human-like stepwise thinking, significantly enhancing performance on complex tasks (Jaech et al. 2024; DeepSeek-AI 2025). A pivotal breakthrough, DeepSeek-R1 (DeepSeek-AI 2025), leverages large-scale RL with formatting and result-oriented rewards to enable LLMs to autonomously generate human-like complex Chain-of-Thought (CoT) reasoning, achieving state-of-the-art results across diverse domains. MLLMs extend these capabilities through systematic integration of cross-modal knowledge representation and task-specific rewards (Yang et al. 2025; Du et al. 2025; Huang et al. 2025; Shen et al.

2026) For multimodal mathematical reasoning, frameworks such as R1-Onevision (Yang et al. 2025), Virgo (Du et al. 2025), and MM-Eureka (Meng et al. 2025) jointly reason over visual and textual inputs to solve quantitative tasks, while Vision-R1 (Huang et al. 2025) combines cold-start initialization with GRPO to refine complex CoT reasoning. For fine-grained classification and grounding tasks, Visual-RFT (Liu et al. 2025b) and R1-Omni (Zhao, Wei, and Bo 2025) utilize GRPO-based reinforcement algorithms and verifiable rewards to enhance contextual reasoning. As for the pixel-level understanding task, SegZero (Liu et al. 2025a) employs reinforcement learning to achieve high-resolution semantic segmentation. Despite these advances, existing MLLMs predominantly focus on *object-centric recognition* rather than unified visual relation understanding, which requires compositional reasoning about geometric, semantic, and functional interactions between objects and contexts. To bridge this gap, Relation-R1 explicitly integrates *relational reasoning* into MLLMs via RL, enabling systematic inference for both scene interpretation and event prediction.

Scene Graph Generation (SGG). SGG is a critical task in scene understanding, with prior approaches falling into two groups: 1) *Two-stage SGG*, which sequentially detects objects and infers pairwise relations but suffers from error propagation (Tang et al. 2020; Zellers et al. 2018; Li et al. 2022, 2023a; Shi et al. 2025); 2) *One-stage SGG*, which unifies detection and relation prediction in end-to-end frameworks (*e.g.*, DETR-based methods (Carion et al. 2020; Li, Zhang, and He 2022; Chen et al. 2020)) but lacks multi-role involved relation modeling. Recent efforts in *open-vocabulary SGG (OVSGG)* employ Vision-Language Models (VLMs) (Chen, Li, and Wang 2024; Li et al. 2023b, 2025; Chen et al. 2024) or MLLMs (Wang et al. 2024; Xu et al. 2025; Li et al. 2024b) to handle novel entities/relations, yet limited modeling of complex multi-entity interactions, and suffer from overfitting in base categories during SFT. Our Relation-R1 addresses OVSGG challenges by adopting a cognitive CoT-guided RL framework to model complex multi-entity interactions and mitigate SFT overfitting, enabling robust zero-shot reasoning without predefined category constraints.

Grounded Situation Recognition (GSR). GSR extends scene understanding by jointly modeling actions (verbs) and their associated semantic roles. Pratt *et al*. (Pratt et al. 2020) pioneered this field with a two-stage RNN-based framework: verb detection followed by noun localization. Subsequent works (Cho et al. 2021; Wei et al. 2022; Cho, Yoon, and Kwak 2022; Cheng et al. 2022) improved this pipeline by integrating transformers and semantic relation modeling, enabling more coherent scene interpretations. Despite these efforts, previous efforts that depend on annotated training samples may face challenges like long-tailed distribution (Tang et al. 2020; Li et al. 2024a; Shao et al. 2024; Chen et al. 2023), potentially limiting real-world applicability. While LEX (Lei et al. 2024) realizes training-free zero-shot GSR, it fails to provide an end-to-end framework and exhibits sub-optimal performance and computational efficiency due to its reliance on LLMs' generated descriptions. This paper proposes Relation-R1, a unified framework that bridges the gap between open-ended capabilities and end-to-end training.

Methodology

Preliminaries. 1) *Binary Relation Detection*: This task requires the model to generate an *open-ended scene graph caption* given an image, explicitly marking semantic relationships between pairwise objects (Wang et al. 2024). The output is a triplet-contained caption annotated with formal tags: objects are denoted by <ref> tags followed by their bounding box coordinates (top-left/bottom-right) in < box>, *e.g.*, <ref>*person*</ref><box>[[x_1, y_1, x_2, y_2]]</box>. Predicates are enclosed in <pred> tags, followed by references to their *subject* and *object* entities via their corresponding bounding boxes. 2) *N-ary Relation Detection:* It aims to output a *open-ended grounded situation frame* (Pratt et al. 2020) structured as a primary activity (*i.e.*, multi-entity involved relation), followed by a sequence of <role> tags along with corresponding bounding box coordinates. Each role explicitly defines the entity's participation in the activity, *e.g.*, <agent>*person*</agent><box>[x_1, y_1, x_2, y_2]</box>.

Supervised Fine-Tuning

As discussed above, directly applying GRPO encounters challenges in generating consistently formatted scene graph captions and grounded situation frame outputs. To address this, we employ SFT as a foundational stage to establish task-specific output format adherence (*c.f.*, Figure 2(a)).

However, relying solely on the target answer format during SFT risks compromising the model's underlying reasoning capabilities by encouraging a superficial alignment with formulated structures (Chu et al. 2025). To mitigate this potential issue, we introduce the **cognitive chain-of-thought** that integrates stepwise cognitive processes such as object classification, object grounding, and visual relationship inference. These cognitive CoTs are explicitly encapsulated within <think> tags, during SFT. This ensures that the model retains its capacity for multi-step reasoning while learning to produce structured outputs conforming to the desired format.

To streamline the think process and prevent the MLLM from overfitting to specific CoT patterns during SFT, we adopt a specific-to-general progressive training strategy:

Template-based CoT. For the specific learning phase, we devise a fixed, template-based CoT. As illustrated in Figure 2(a), this template comprises a predefined sequence of steps for both binary and *N*-ary relation detection. To elaborate, for binary relation detection, the template includes: 1) Object Existence (*e.g.*, `the objects present ...`); 2) Object Localization (*e.g.*, `the objects are located ...`); and 3) Relation Existence (*e.g.*, `the relations present ...`). For *N*-ary relation detection, it involves: 1) Activity Recognition (*e.g.*, `the primary activity is ...`); 2) Entities and Roles Recognition (*e.g.*, `the entities engaged in the activity are...`); and 3) Entity Localization (*e.g.*, `the entities are located in ...`).

MLLM-generated CoT. For the general reasoning process, we design a **cognitive CoT generation prompt** that explicitly makes an extra strong MLLM (*e.g.*, Qwen 2.5-VL (Bai et al. 2025)) to generate step-by-step reasoning processes. To be specific, this cognitive CoT generation prompt consists of a task definition, a ground-truth (GT) scene graph

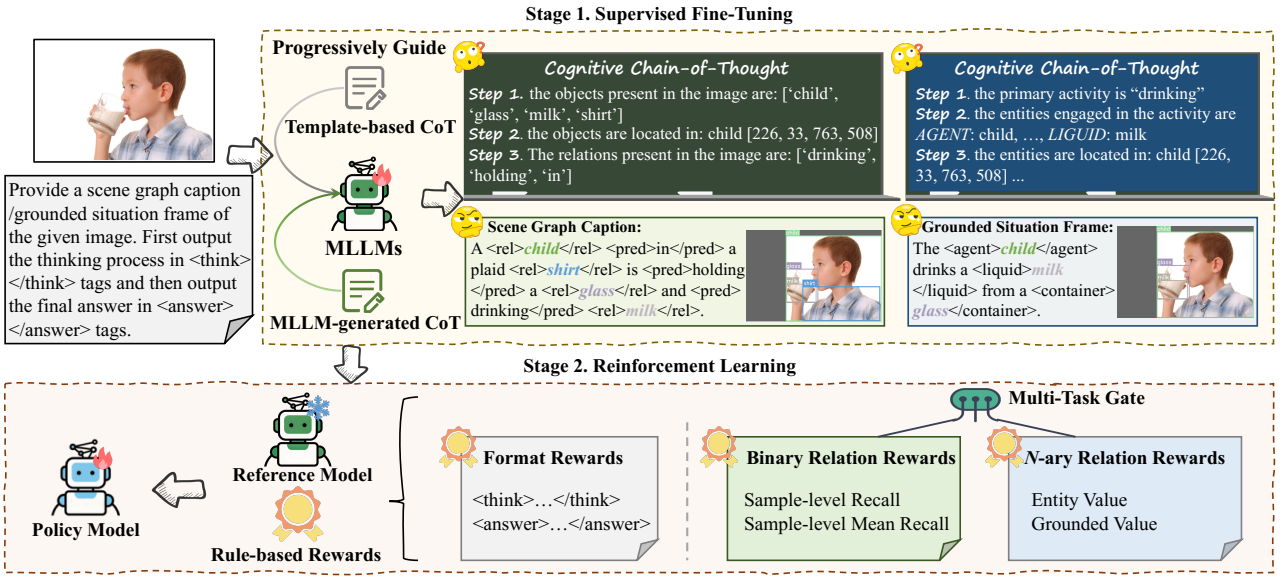

Figure 2: **Relation-R1** framework. (a) **SFT** with progressive guidance of cognitive CoT, which enforces task-specific reasoning processes and format alignment. (b) **RL** with GRPO, employs rule-based rewards (*e.g.*, format rewards, binary relation rewards, and N-ary relation rewards) to cultivate robust relation reasoning and generalization ability through policy gradient updates.

of the current image, and a CoT generation instruction. The former one makes the MLLM generate the correct task format. Meanwhile, the latter two anchors reasoning content to visual evidence, ensuring that the MLLM's cognitive CoT aligns with both semantic validity and visual grounding rather than relying on category priors. The detailed cognitive CoT generation prompt is left in the Appendix.

Progressive CoT Guidance. With a progressive paradigm that involves initial SFT training on easy-to-learn template-based CoTs and subsequent fine-tuning on a small number of MLLM-generated CoTs, the MLLM gains foundational capability to generate format-compliant outputs and grounded reasoning processes, contributing to GRPO training.

Group Relative Policy Optimization

Following the Deepseek-R1 (DeepSeek-AI 2025), the group relative policy optimization is adopted as the RL algorithm for optimizing our Relation-R1 framework, as illustrated in Figure 2(b). Distinct from critic-dependent methods (Schulman et al. 2017), GRPO directly calculates advantage estimates through response group comparisons sampled from the policy model, inherently eliminating the need for an extra critic network and reducing computational complexity (DeepSeek-AI 2025). Key components include the a question q, policy model π_θ, response group $\{o_1, o_2, \ldots, o_G\}$, and a frozen reference model $\pi_{\theta_{\text{ref}}}$, which acts as a regularization baseline to preserve prior knowledge. The GRPO objective function $J_{\text{GRPO}}(\theta)$ balances reward maximization and policy stability by leveraging the following objectives:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}} \left[\frac{1}{G}\sum_{i=1}^{G}\min\left(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)A_i\right) - \beta D_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\right], \tag{1}$$

where $\rho_i = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}$ quantifies policy change, ϵ controls clipping thresholds, and β penalizes deviations via the KL divergence term. The advantage score A_i is standardized as:

$$A_i = \frac{r_i - \text{mean}\left(\{r_1, \ldots, r_G\}\right)}{\text{std}\left(\{r_1, \ldots, r_G\}\right)}, \tag{2}$$

here r_i denotes the reward for response o_i. The KL divergence term enforces proximity to the reference policy:

$$D_{KL}(\pi_\theta\|\pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(o_i|q)}{\pi_\theta(o_i|q)} - \log\left(\frac{\pi_{\text{ref}}(o_i|q)}{\pi_\theta(o_i|q)}\right) - 1. \tag{3}$$

This term ensures controlled exploration without excessive divergence from $\pi_{\theta_{\text{ref}}}$.

The reward r_i is formulated to align with the model's learning objective for the unified relation understanding. It incorporates rule-based rewards, which consist of three crucial components: Format compliance rewards, binary relation rewards, and N-ary relation rewards. The former enforces syntactic validity by penalizing outputs that deviate from predefined structural constraints. The latter two components focus on task-specific performance.

Format Rewards. This reward enforces strict adherence to the reasoning template format: `<think>` *thinking process* `</think><answer>` *answer* `</answer>`. The reward function is defined as:

$$r_{\text{form}}(o_i) = \begin{cases} 1 & \text{if } o_i \text{ adheres to the format,} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

This binary mechanism ensures syntactic compliance while allowing content flexibility.

Binary Relation Rewards. Inspired by established evaluation practices (Yang et al. 2022; Wang et al. 2024), we devise the sample-level triplet recall R and sample-level mean recall

mR in the response of each sample. The former is the ratio of correctly predicted triplets to the total ground-truth triplets in a sample; the latter is the average recall across all predicate categories. Specifically, a triplet is deemed correct if:

- **Triplet Accuracy**: All entity and relation categories (`subject`, `predicate`, `object`) in this sample are correctly identified.
- **Spatial Localization**: The predicted bounding boxes for the subject and object achieve an Intersection over Union (IoU) >= 0.5 with their respective ground-truth boxes.

The final binary relation reward is computed as a weighted combination of these two metrics:

$$r_{\text{binary}}(o) = \alpha \cdot R + (1 - \alpha) \cdot mR, \tag{5}$$

where α balances the trade-off between two metrics.

***N*-ary Relation Rewards.** Similarly, to test higher-order relational structures in grounded situation frames, we devise *N*-ary relation rewards to assess the model's ability to detect primary activity (verb) and multi-entity interactions. Specifically, we calculate the entity value V_e and grounded value V_{grnd} for each sample (Pratt et al. 2020). They represent the correct proportion in predicted entity classes and roles, and the spatial localization of those entities, respectively:

- **Entity Accuracy**: An entity within an event is considered correct if its predicted category and semantic role align with the GT, and the predicted activity matches the GT.
- **Spatial Localization**: The predicted bounding box of each role obtains IoU >=0.5 with the GT's bounding box.

The eventual *N*-ary relation reward is computed as a weighted combination of these two accuracies:

$$r_{\text{n-ary}}(o) = \beta \cdot V_e + (1 - \beta) \cdot V_{grnd}, \tag{6}$$

where the hyperparameter β balances the trade-off between the two values. As the reward's calculation depends on accurate verb prediction, they jointly evaluate primary activity and multi-entity interactions. During training, we design a multi-task gate that dynamically selects task-specific rewards based on the presence of <ref> tags in the solutions.

Experiment

Datasets. 1) **Panoptic Scene Graph (PSG)**: The official PSG dataset (Yang et al. 2022) comprises 48,749 annotated images of which 46,563 are for training and 2,186 are for testing. It integrates 80 "thing" object categories and 53 "stuff", alongside 56 relation categories. Notably, ASMv2 (Wang et al. 2024) reconstructs the dataset into a scene graph caption format, utilizing a split of 42,250 training images and 1,000 evaluation images. To mark fair comparison, we followed ASMv2 (Wang et al. 2024), using scene graph caption where objects are tagged with <ref> and paired with bounding box coordinates specifying their locations, while predicates are marked with <pred> and linked to two bounding boxes indicating their subject and object entities for training. Besides, we also present results using the conventional scene graph format, *i.e.*, a list of visual triplets structured as [subject, subject bounding box, object, object bounding box, relation] on the official dataset[1]. 2) **SWiG**: The SWiG dataset (Pratt et al. 2020) extends the imSitu dataset (Yatskar, Zettlemoyer, and Farhadi 2016) by adding extra bounding box annotations for semantic roles, with 69.3% of semantic roles localized. Each image is annotated with a verb and a variable number of semantic roles (1∼6 roles per image), where each verb is associated with three verb frames annotated by independent human annotators. The dataset comprises 25,200 testing images spanning 504 verb categories, 190 semantic role categories, and 9,929 noun entity categories. Unlike traditional output structures that merely predict categories and bounding boxes, we constructed grounded situation frames by filling verb-centric templates (*e.g.*, "The AGENT drinks a LIQUID from a CONTAINER") with three critical components: role tags (*e.g.*, <agent>, entity categories (*e.g.*, child), and coordinates of corresponding bounding boxes in <box> tags.

[1]More details of the scene graph format are in the Appendix.

Method	Size	Recall	mRecall	Mean
Close-ended				
IMP (Xu et al. 2017)	-	16.50	6.50	11.50
MOTIFS (Zellers et al. 2018)	-	20.00	9.10	14.55
VCTree (Tang et al. 2019)	-	20.60	9.70	15.15
GPSNet (Lin et al. 2020)	-	17.80	7.00	12.4
PSGFormer (Yang et al. 2022)	-	18.60	16.70	17.65
Open-ended				
TextPSG* (Zhao et al. 2023)	-	4.80	-	-
R1-SGG* (Chen et al. 2025)	2B	**27.83**	17.03	22.43
R1-SGG* (Chen et al. 2025)	7B	**28.77**	17.55	23.16
Relation-R1*(Ours)	3B	25.87	**21.32**	**23.60**
ASMv2† (Wang et al. 2024)	13B	14.20	10.30	12.23
SpaceSGG† (Xu et al. 2025)	13B	15.43	13.23	14.33
Relation-R1† (Ours)	3B	**22.33**	**20.07**	**21.20**

Table 1: Comparison with various SGG models on PSG dataset (Yang et al. 2022). † denotes training/test with *scene graph **caption** format* following ASMv2 (Wang et al. 2024). $\star$ indicates training/test in the *standard scene graph format*[1].

Evaluation Metrics. For *binary relation detection*, we followed the standard SGG evaluation protocol (Tang et al. 2020; Wang et al. 2024) 1) **Recall**: The ratio of correctly predicted triplets to the total number of ground-truth triplets, with both label accuracy and bounding box IoU >= 0.5 for subject and object. 2) **mean Recall** (**mRecall**): The average recall across all predicate classes. For *N-ary relation detection*, we adopted the same evaluation metrics as (Pratt et al. 2020): 1) **verb**: the accuracy of verb prediction. 2) **value**: the noun accuracy for individual semantic role. 3) **value-all (val-all)**: Overall noun prediction accuracy across all semantic roles in an event. 4) **grounded-value (grnd)**: Bounding box accuracy for each semantic role, requiring an IoU >= 0.5 between predicted and ground-truth boxes. 5) **grounded-value-all (grnd-all)**: entire bounding box accuracy across all semantic roles in an event.

Implementation Details. Refer to the Appendix.

Comparison with State-of-the-Art Methods

We compared Relation-R1 for unified relation comprehension with the existing state-of-the-art binary relation detection (Table 1) and *N*-ary relation detection approaches in (Table 2).

Binary Relation Detection. For the open-ended scene graph

Method		Answer	Verb	Value	Value-all	Grnd	Grnd-all
ISL (Pratt et al. 2020)	ECCV'20	Close-ended	39.36	30.09	18.62	22.73	7.72
JSL (Pratt et al. 2020)	ECCV'20		39.94	31.44	18.87	24.86	9.66
GSRTR (Cho et al. 2021)	BMVC'21		40.63	32.15	19.28	25.49	10.10
CoFormer (Cho, Yoon, and Kwak 2022)	CVPR'22		44.66	35.98	22.22	29.05	12.21
SituFormer (Wei et al. 2022)	AAAI'22		44.20	35.24	21.86	29.22	13.41
GSRFormer (Cheng et al. 2022)	ACM MM'22		46.53	37 .48	23.32	31.53	14.23
OpenSU (Liu et al. 2023b)	ICCVW'23	Open-ended	50.10	41.20	26.56	34.27	15.70
Relation-R1 (Ours)			**57.26**	**46.66**	**30.92**	**40.21**	**30.18**

Table 2: Performance(%) comparison with various GSR models on the SWiG (Pratt et al. 2020) dataset.

Method	CoT	Binary Relation		N-ary Relation				
		Recall	mRecall	Verb	Value	Value-all	Grnd	Grnd-all
SFT	-	14.83	13.86	56.64	42.65 (60.45)	22.11 (25.21)	37.70 (54.05)	16.35 (19.02)
SFT + RL	Template-based	20.24	17.31	58.38	47.75 (66.74)	32.00 (35.67)	41.27 (55.88)	31.16 (38.29)
SFT + RL	MLLM-generated	20.66	20.30	53.00	42.27 (63.86)	26.67 (30.79)	35.69 (52.43)	25.89 (33.95)
SFT + RL	Progressive	22.57	20.57	71.04	61.26 (78.19)	44.98 (49.73)	49.43 (61.25)	36.09 (42.35)

Table 3: Performance comparison (%) across various cognitive CoT strategies trained on single binary relation detection and N-ary relation detection, separately. **Blue** indicates metrics without correct verb constraints.

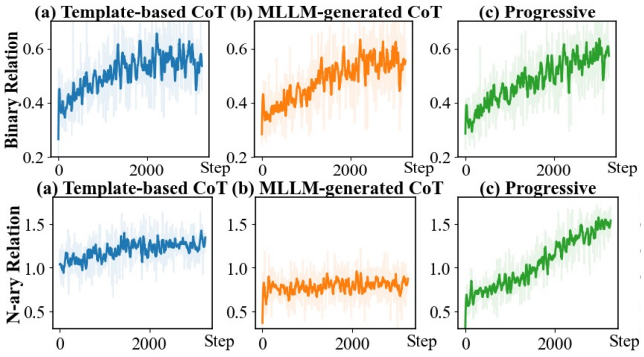

Figure 3: The statistics of binary and N-ary relation rewards.

caption (*c.f.* Table 1[2]), Relation-R1 achieves state-of-the-art performance with the highest Recall of **22.33**%, mRecall of **20.07**% and overall Mean of **21.20**%. This substantially surpasses prior open-ended methods such as LLaVA-SpaceSGG (Xu et al. 2025) (15.43% Recall, 13.23% mRecall, 14.33% Mean). Notably, even when evaluated on the standard scene graph format[1], which presents a less challenging scenario than scene graph captioning due to the reduced need for contextual inference, Relation-R1 leads across all evaluation metrics. Furthermore, Relation-R1 achieves these results with a significantly smaller 3B parameter model, in contrast to the larger 13B parameter models employed by recent competitors ASMv2 (Wang et al. 2024) and LLaVA-SpaceSGG, demonstrating superior parameter efficiency.

***N*-ary Relation Detection.** Within the open-ended grounded situation frame setting (*c.f.* Table 2[2]), Relation-R1 also establishes a new SOTA, achieving top performance across all metrics: Verb (**57.26**%), Value (**46.66**%), Value-all (**30.92**%), Grnd (**40.21**%), and Grnd-all (**30.18**%). Notably, Relation-R1 significantly surpasses the previous best open-ended model, OpenSU (Liu et al. 2023b), exhibiting substantial gains, particularly with a remarkable **+14.48**% absolute improvement on the challenging Grnd-all metric. This clearly demonstrates Relation-R1's superior capacity for comprehending and grounding complex N-ary relationships in visual scenes.

[2]This Relation-R1 version is guided by the template-based CoTs.

Diagnostic Experiment

To gain more insights and reduce the mutual influence between the two tasks during training, we conducted a set of comprehensive studies on the Relation-R1 trained with binary and N-ary relation detection, respectively.

Cognitive Chain-of-Thoughts. Table 3 compares different cognitive CoT strategies for binary and N-ary relation detection. As seen, introducing RL with various CoT strategies (SFT + RL) demonstrates clear benefits. The template-based CoT improves binary relation Recall to **20.24**% and mRecall to **17.31**%. For N-ary relations, it enhances performance across the board compared to the constrained SFT baseline, for instance, achieving **47.75**% on Value and **31.16**% on Grnd-all. The MLLM-generated CoT slightly outperforms the template-based approach in binary relation detection, but underperforms in N-ary relation detection. This is because N-ary relation detection metrics are often limited by verb accuracy, the model may generate a verb different from the GT yet still semantically plausible. Notably, when the verb constraint is removed, MLLM-generated CoT yields significant gains, particularly in Grnd-all (**+14.93**%). While template-based and MLLM-generated CoTs offer improvements over SFT, the progressive strategy stands out. This approach significantly outperforms all others across all metrics, achieving **22.57**% Recall in binary relation detection and top scores (*e.g.*, **71.04**% Verb accuracy) in N-ary relation detection. These results highlight the superior ability of the progressive CoT to guide complex relation detection.

Reward Analysis. We visualized the reward variation curves for binary and N-ary relation detection in Figure 3, respectively. It can be seen that at the beginning stage, both tasks

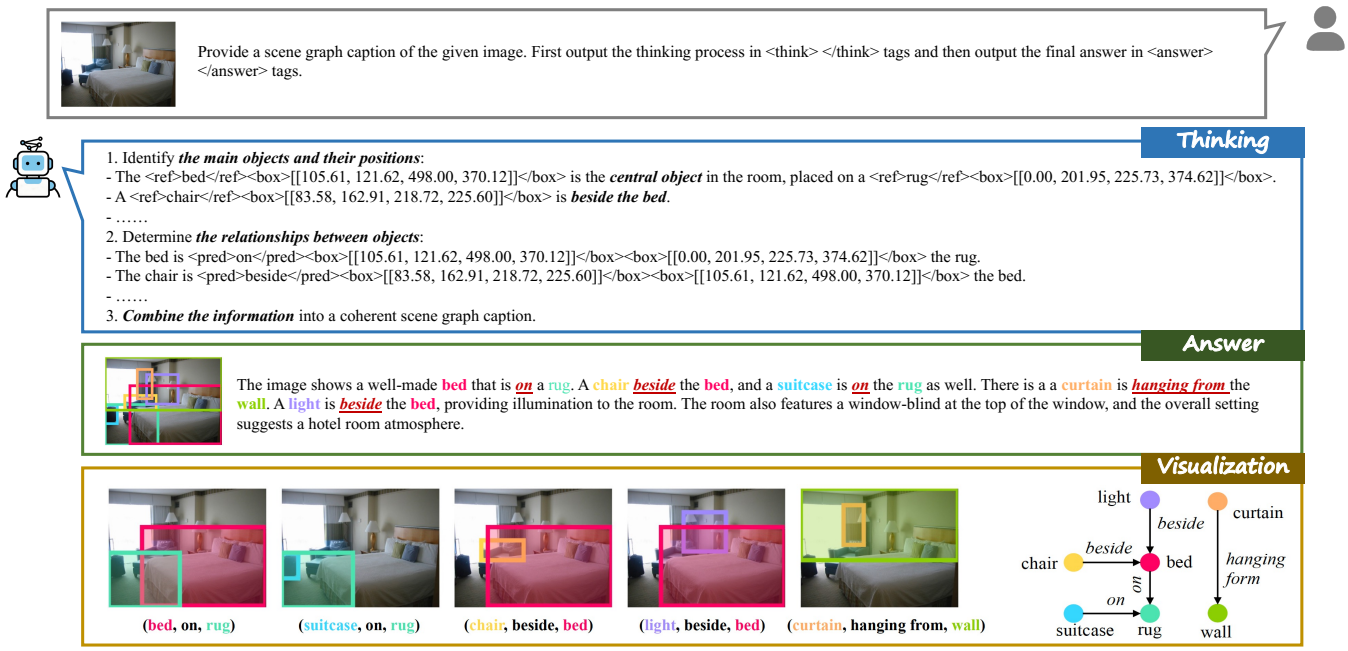

Figure 4: Qualitative results of binary relation detection.

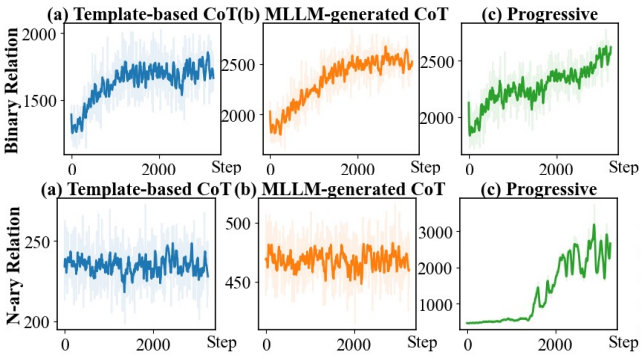

Figure 5: Completion lengths of binary and N-ary relation.

obtains decent rewards due to the foundation laid by SFT. With the gradual progress of GRPO, the rewards of both tasks show a gradually increasing trend. In binary relation detection, guided with progressive CoT yields a slightly enhanced overall reward trajectory, although the improvement is not significant. This might be since the gap between the CoT generated by MLLM (*e.g.*, Qwen (Bai et al. 2025)) and that produced by the template is not obvious. Conversely, N-ary relation detection exhibits a significant surge in reward, primarily attributed to the model's emergent capability to generate synonymous grounded situation frames.

Completion Length Analysis. The statistic of completion length during GRPO training is in Figure 5. Consistent with the reward trends, the completion lengths for binary relation detection exhibit a gradual increase as GRPO progresses, suggesting the generation of more detailed and descriptive outputs over time. Notably, the template-based CoT strategy yields significantly shorter completions compared to the more flexible MLLM-generated and progressive CoT strategies, since only the cognitive process is involved. The progressive CoT guidance, by enabling the model to explore synonymous relational expressions and incorporate more nuanced details, leads to an increase in the average completion length.

Qualitative Results. Figure 4 shows a qualitative example of binary relation detection. As seen, the visualized thinking process reveals how Relation-R1 effectively decomposes the complex task into constituent cognitive steps: object classification, localization, and relation inference. For instance, after identifying the central `bed`, Relation-R1 pinpoints related objects *e.g.*, `rug` and `chair`, and explicitly grounds the inferred spatial relations (`on`, `beside`) between them. Our approach demonstrates the successful decomposition of the scene into accurate, grounded object-relation pairs.

Conclusion

This paper reveals key limitations in MLLM visual relation understanding, particularly for N-ary relations and language-induced bias. We propose Relation-R1, a unified two-stage framework: build a reasoning foundation and structured outputs via cognitive CoT-guided SFT, followed by GRPO, promoting visual-semantic grounding and robust generalization. Besides, progressively applied cognitive CoT guidance, transitioning from template-based to MLLM-generated reasoning, further enhances generalization, especially for synonymous N-ary relations. Experiments on PSG and SWiG validate Relation-R1's effectiveness in both binary and N-ary relation detection. Employing progressive CoT-guided SFT with R1-inspired refinement, Relation-R1 advances MLLMs towards more human-like relational reasoning.

Acknowledgments

This work was supported by the National Natural Science Foundation of China Young Scholar Fund (62402408), the Hong Kong SAR RGC General Research Fund under Grant (16208823), and the Hong Kong SAR RGC Early Career Scheme (26208924). This research was partially conducted by ACCESS – AI Chip Center for Emerging Smart Systems, supported by the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government. This work was also supported by the Hong Kong SAR RGC General Research Fund under Grant (16208823) and the Hong Kong SAR RGC Early Career Scheme (26208924).

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*.

Chen, G.; Li, J.; and Wang, W. 2024. Scene graph generation with role-playing large language models. *NeurIPS*.

Chen, G.; Li, L.; Luo, Y.; and Xiao, J. 2023. Addressing Predicate Overlap in Scene Graph Generation with Semantic Granularity Controller. In *ICME*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607.

Chen, Z.; Wu, J.; Lei, Z.; Pollefeys, M.; and Chen, C. W. 2025. Compile Scene Graphs with Reinforcement Learning. *arXiv preprint arXiv:2504.13617*.

Chen, Z.; Wu, J.; Lei, Z.; Zhang, Z.; and Chen, C. 2024. Expanding Scene Graph Boundaries: Fully Open-vocabulary Scene Graph Generation via Visual-Concept Alignment and Retention. In *ECCV*.

Cheng, Z.-Q.; Dai, Q.; Li, S.; Mitamura, T.; and Hauptmann, A. 2022. Gsrformer: Grounded situation recognition transformer with alternate semantic attention refinement. In *ACM MM*, 3272–3281.

Cho, J.; Yoon, Y.; and Kwak, S. 2022. Collaborative Transformers for Grounded Situation Recognition. In *CVPR*, 19659–19668.

Cho, J.; Yoon, Y.; Lee, H.; and Kwak, S. 2021. Grounded Situation Recognition with Transformers. In *BMVC*.

Chu, T.; Zhai, Y.; Yang, J.; Tong, S.; Xie, S.; Schuurmans, D.; Le, Q. V.; Levine, S.; and Ma, Y. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Du, Y.; Liu, Z.; Li, Y.; Zhao, W. X.; Huo, Y.; Wang, B.; Chen, W.; Liu, Z.; Wang, Z.; and Wen, J.-R. 2025. Virgo: A Preliminary Exploration on Reproducing o1-like MLLM. *arXiv preprint arXiv:2501.01904*.

Huang, W.; Jia, B.; Zhai, Z.; Cao, S.; Ye, Z.; Zhao, F.; Hu, Y.; and Lin, S. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *CVPR*, 9579–9589.

Lei, J.; Li, L.; Wang, C.; Xiao, J.; and Chen, L. 2024. Seeing Beyond Classes: Zero-Shot Grounded Situation Recognition via Language Explainer. In *ACM MM*.

Li, L.; Chen, G.; Xiao, J.; Yang, Y.; Wang, C.; and Chen, L. 2023a. Compositional feature augmentation for unbiased scene graph generation. In *ICCV*, 21685–21695.

Li, L.; Chen, L.; Huang, Y.; Zhang, Z.; Zhang, S.; and Xiao, J. 2022. The devil is in the labels: Noisy label correction for robust scene graph generation. In *CVPR*, 18869–18878.

Li, L.; Xiao, J.; Chen, G.; Shao, J.; Zhuang, Y.; and Chen, L. 2023b. Zero-shot visual relation detection via composite visual cues from large language models. In *NeurIPS*.

Li, L.; Xiao, J.; Shi, H.; Zhang, H.; Yang, Y.; Liu, W.; and Chen, L. 2024a. Nicest: Noisy label correction and training for robust scene graph generation. *IEEE TPAMI*.

Li, L.; Zhang, C.; Zhang, D.; Sun, C.; Li, C.; and Chen, L. 2025. Interaction-Centric Knowledge Infusion and Transfer for Open Vocabulary Scene Graph Generation. *Advances in Neural Information Processing Systems*.

Li, R.; Zhang, S.; and He, X. 2022. Sgtr: End-to-end scene graph generation with transformer. In *CVPR*, 19486–19496.

Li, R.; Zhang, S.; Lin, D.; Chen, K.; and He, X. 2024b. From Pixels to Graphs: Open-Vocabulary Scene Graph Generation with Vision-Language Models. In *CVPR*, 28076–28086.

Lin, X.; Ding, C.; Zeng, J.; and Tao, D. 2020. Gps-net: Graph property sensing network for scene graph generation. In *CVPR*, 3746–3753.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. *NeurIPS*, 36: 34892–34916.

Liu, R.; Zhang, J.; Peng, K.; Zheng, J.; Cao, K.; Chen, Y.; Yang, K.; and Stiefelhagen, R. 2023b. Open Scene Understanding: Grounded Situation Recognition Meets Segment Anything for Helping People with Visual Impairments. In *ICCVW*.

Liu, Y.; Peng, B.; Zhong, Z.; Yue, Z.; Lu, F.; Yu, B.; and Jia, J. 2025a. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*.

Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025b. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

Meng, F.; Du, L.; Liu, Z.; Zhou, Z.; Lu, Q.; Fu, D.; Shi, B.; Wang, W.; He, J.; Zhang, K.; et al. 2025. MM-Eureka: Exploring Visual Aha Moment with Rule-based Large-scale Reinforcement Learning. *arXiv preprint arXiv:2503.07365*.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2024. Kosmos-2: Grounding Multimodal Large Language Models to the World. In *ICLR*.

Pratt, S.; Yatskar, M.; Weihs, L.; Farhadi, A.; and Kembhavi, A. 2020. Grounded situation recognition. In *ECCV*, 314–332.

Rasheed, H.; Maaz, M.; Shaji, S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2024. Glamm: Pixel grounding large multimodal model. In *CVPR*, 13009–13018.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shao, F.; Luo, Y.; Gao, F.; Yang, Y.; and Xiao, J. 2024. Knowledge-guided causal intervention for weakly-supervised object localization. *IEEE Transactions on Knowledge and Data Engineering*, 36(11): 6477–6489.

Shen, K.; Quan, R.; Miao, J.; Xiao, J.; and Yang, Y. 2026. Tarpro: Targeted protection against malicious image editing. In *AAAI*.

Shi, H.; Li, L.; Xiao, J.; Zhuang, Y.; and Chen, L. 2025. From easy to hard: Learning curricular shape-aware features for robust panoptic scene graph generation. *IJCV*, 133(1): 489–508.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased scene graph generation from biased training. In *CVPR*, 3716–3725.

Tang, K.; Zhang, H.; Wu, B.; Luo, W.; and Liu, W. 2019. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, 6619–6628.

Wang, W.; Ren, Y.; Luo, H.; Li, T.; Yan, C.; Chen, Z.; Wang, W.; Li, Q.; Lu, L.; Zhu, X.; et al. 2024. The all-seeing project v2: Towards general relation comprehension of the open world. In *ECCV*, 471–490. Springer.

Wei, M.; Chen, L.; Ji, W.; Yue, X.; and Chua, T.-S. 2022. Rethinking the Two-Stage Framework for Grounded Situation Recognition. In *AAAI*, 2651–2658.

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene graph generation by iterative message passing. In *CVPR*, 5410–5419.

Xu, M.; Wu, M.; Zhao, Y.; Li, J. C. L.; and Ou, W. 2025. LLaVA-SpaceSGG: Visual Instruct Tuning for Open-vocabulary Scene Graph Generation with Enhanced Spatial Relations. In *WACV*.

Yang, J.; Ang, Y. Z.; Guo, Z.; Zhou, K.; Zhang, W.; and Liu, Z. 2022. Panoptic scene graph generation. In *ECCV*, 178–196. Springer.

Yang, Y.; He, X.; Pan, H.; Jiang, X.; Deng, Y.; Yang, X.; Lu, H.; Yin, D.; Rao, F.; Zhu, M.; Zhang, B.; and Chen, W. 2025. R1-Onevision: Advancing Generalized Multimodal Reasoning through Cross-Modal Formalization. *arXiv preprint arXiv:2503.10615*.

Yatskar, M.; Zettlemoyer, L.; and Farhadi, A. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *CVPR*, 5534–5542.

You, H.; Zhang, H.; Gan, Z.; Du, X.; Zhang, B.; Wang, Z.; Cao, L.; Chang, S.-F.; and Yang, Y. 2023. Ferret: Refer and Ground Anything Anywhere at Any Granularity. In *ICLR*.

Yuan, Z.; Deng, J.; Ming, R.; Lang, F.; and Yang, X. 2024. SR-LIVO: LiDAR-Inertial-Visual Odometry and Mapping With Sweep Reconstruction. *IEEE Robotics and Automation Letters*, 9(6): 5110–5117.

Yuan, Z.; Lang, F.; Deng, J.; Luo, H.; and Yang, X. 2025. Voxel-SVIO: Stereo Visual-Inertial Odometry based on Voxel Map. *IEEE Robotics and Automation Letters*, 10(6): 6352–6359.

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural motifs: Scene graph parsing with global context. In *CVPR*, 5831–5840.

Zhang, H.; Li, H.; Li, F.; Ren, T.; Zou, X.; Liu, S.; Huang, S.; Gao, J.; Leizhang; Li, C.; et al. 2024. Llava-grounding: Grounded visual chat with large multimodal models. In *ECCV*, 19–35. Springer.

Zhao, C.; Shen, Y.; Chen, Z.; Ding, M.; and Gan, C. 2023. TextPSG: Panoptic Scene Graph Generation from Textual Descriptions. In *ICCV*, 2839–2850.

Zhao, J.; Wei, X.; and Bo, L. 2025. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv e-prints*, arXiv–2503.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *ICLR*.