GeM-VG: Towards Generalized Multi-image Visual Grounding with Multimodal Large Language Models

Shurong Zheng[1, 2, 3], **Yousong Zhu**[4], **Hongyin Zhao**[1], **Fan Yang**[1, 2, 3], **Yufei Zhan**[1, 3], **Ming Tang**[1, 3], **Jinqiao Wang**[1, 2, 3, 5*],

[1] Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[4]School of Artificial Intelligence, China University of Mining and Technology-Beijing, Beijing, China
[5]Wuhan AI Research, Wuhan, China
zhengshurong2023@ia.ac.cn, yousong.zhu@cumtb.edu.cn, zhaohongyin2020@ia.ac.cn, yangfan_2022@ia.ac.cn, zhanyufei2021@ia.ac.cn, tangm@nlpr.ia.ac.cn, jqwang@nlpr.ia.ac.cn

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated impressive progress in single-image grounding and general multi-image understanding. Recently, some methods begin to address multi-image grounding. However, they are constrained by single-target localization and limited types of practical tasks, due to the lack of unified modeling for generalized grounding tasks. Therefore, we propose GeM-VG, an MLLM capable of **Ge**neralized **M**ulti-image **V**isual **G**rounding. To support this, we systematically categorize and organize existing multi-image grounding tasks according to their reliance of cross-image cues and reasoning, and introduce the MG-Data-240K dataset, addressing the limitations of existing datasets regarding target quantity and image relation. To tackle the challenges of robustly handling diverse multi-image grounding tasks, we further propose a hybrid reinforcement finetuning strategy that integrates chain-of-thought (CoT) reasoning and direct answering, considering their complementary strengths. This strategy adopts an R1-like algorithm guided by a carefully designed rule-based reward, effectively enhancing the model's overall perception and reasoning capabilities. Extensive experiments demonstrate the superior generalized grounding capabilities of our model. For multi-image grounding, it outperforms the previous leading MLLMs by 2.0% and 9.7% on MIG-Bench and MC-Bench, respectively. In single-image grounding, it achieves a 9.1% improvement over the base model on ODINW. Furthermore, our model retains strong capabilities in general multi-image understanding.

Introduction

Traditional visual grounding encompasses tasks such as referring expression comprehension (Yu et al. 2016; Nagaraja, Morariu, and Davis 2016), phrase grounding (Plummer et al. 2015), and object detection (Lin et al. 2014), which involve localizing target regions within a single image based on simple textual descriptions. With the advancements in MLLMs, some works (Chen et al. 2023; Zhan et al. 2024a; Peng et al. 2023; You et al. 2023; Zhan et al. 2024b) leverage the powerful multimodal comprehension capabilities of MLLMs to facilitate visual grounding tasks. Despite their effectiveness, these works are limited to single-image scenarios.

The abilities of multi-image perception and reasoning are crucial for real-world applications such as autonomous driving and GUI agents, which depend on contextual cues across multiple images. While some works (Li et al. 2024b; Jiang et al. 2024; Li et al. 2024a; Ye et al. 2024) investigate image-level tasks in multi-image scenarios, region-level compre-

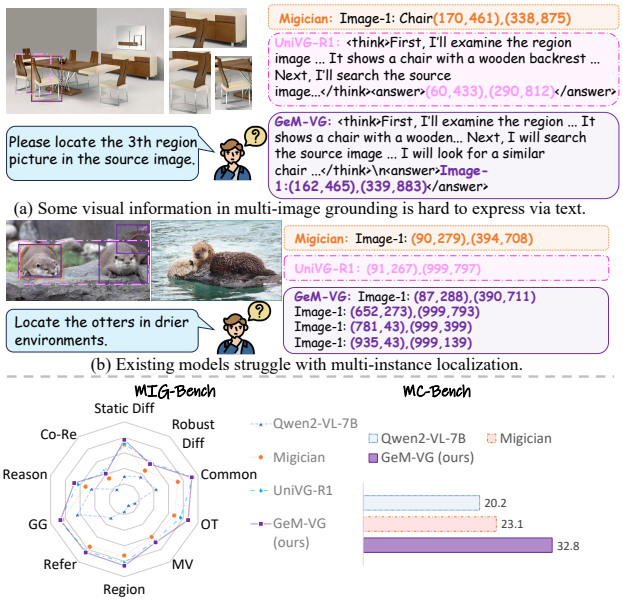


Figure 1: **(Top)** The limitations of existing multi-image grounding models: (a) The reasoning-guided model struggles with fine-grained perception tasks, where the visual cues are hard to verbalize. (b) They fail to localize all referring instances in multi-object scenarios. **(Bottom)** Our model, GeM-VG, achieves strong performance across a range of multi-image grounding tasks.

*Corresponding author.

hension remains relatively underexplored. Recently, Migician (Li et al. 2025) introduces a benchmark comprising diverse multi-image grounding tasks, laying a foundation for this emerging direction. UniVG-R1 (Bai et al. 2025b) further enhances reasoning via reinforcement learning (RL). However, these works focus predominantly on single-target grounding. MC-Bench (Xu, Zhu, and Yang 2024) introduces a more challenging multi-image grounding task with multiple targets, revealing existing MLLMs' limitations in real-world scenarios.

In this work, we aim to advance generalized visual grounding in multi-image scenarios, and propose GeM-VG, an MLLM for **Ge**neralized **M**ulti-image **V**isual **G**rounding. To cover comprehensive multi-image grounding scenarios, we systematically categorize existing multi-image grounding tasks based on cognitive requirements and image relations. Due to the limited target quantity and task scenarios in previous multi-image grounding datasets (Li et al. 2025; Bai et al. 2025b), we curate a multi-image grounding dataset MG-Data-240K to support broader scenarios. As multi-image grounding involves increasingly complex multimodal contexts, it requires more sophisticated perception and reasoning. Inspired by the recent success of reasoning models (Jaech et al. 2024; Guo et al. 2025), we design a new rule-based reward and adapt the R1-like reinforcement learning method to generalized grounding tasks. Due to the lack of basic multi-image grounding capability in the base model, we adopt a progressive three-stage training strategy: (1) supervised finetuning with short-answer data to build fundamental grounding capabilities, (2) CoT-based supervised finetuning to guide reasoning, and (3) GRPO training guided by multi-dimensional visual feedback to further enhance localization and reasoning.

Through supervised fine-tuning with data in different answering modes, we find that CoT reasoning does not consistently outperform direct-answer. In tasks relying on detailed perceptual cues, directly predicting locations avoids the ambiguity of verbalizing abstract visual cues (e.g., Figure 1). Moreover, CoT provides limited benefits in tasks with minimal reasoning demands. To leverage the complementary strengths of both modes, we propose a hybrid finetuning strategy that encourages exploration and optimization of both answering modes during training.

We evaluate GeM-VG on two multi-image grounding benchmarks: MIG-Bench (Li et al. 2025) and MC-Bench (Xu, Zhu, and Yang 2024). On MIG-Bench, GeM-VG surpasses the previous state-of-the-art by 2.0%. On MC-Bench, it outperforms Migician by 9.7% and the base model Qwen2VL-7B by 12.6%. Additionally, our model achieves consistent improvements on single-image grounding and general multimodal understanding tasks. The main contributions of this work are summarized as follows:

1. We provide a systematic taxonomy of existing multi-image grounding tasks and introduce the MG-Data-240K dataset, which encompasses diverse tasks and varying numbers of targets.
2. We propose GeM-VG, an MLLM for generalized multi-image grounding, along with a hybrid RL finetuning strategy that integrates both chain-of-thought and direct-answer modes, tailored for grounding tasks with arbitrary numbers of targets.
3. Extensive experimental results demonstrate that our method consistently improves the performance across various multi-image grounding tasks, while maintaining strong generalization capabilities.

Related Work

Multimodal Large Language Models for Visual Grounding

Recent advances in MLLMs have led to impressive progress in general vision-language understanding and reasoning. By leveraging the powerful multimodal comprehension capabilities of MLLMs, visual grounding tasks are also innovated. Some works (Chen et al. 2023; Zhan et al. 2024a; You et al. 2023) focus on enabling MLLMs to support a range of grounding tasks. With improved reasoning capacity, some studies (Liu et al. 2025; Ma et al. 2025) explore complex reasoning grounding. However, these works are limited to single-image scenarios, hindering the applicability in broader real-world scenarios. Recently, Migician (Li et al. 2025) introduces a multi-image grounding benchmark and enables the model to perform free-form multi-image grounding via supervised finetuning. UniVG-R1 (Bai et al. 2025b) further enhances reasoning via reinforcement learning. Nevertheless, they primarily focus on single-object grounding and struggle with multi-instance cases. In contrast, our work aims to address generalized multi-image grounding tasks.

Vision-Language Reinforcement Learning

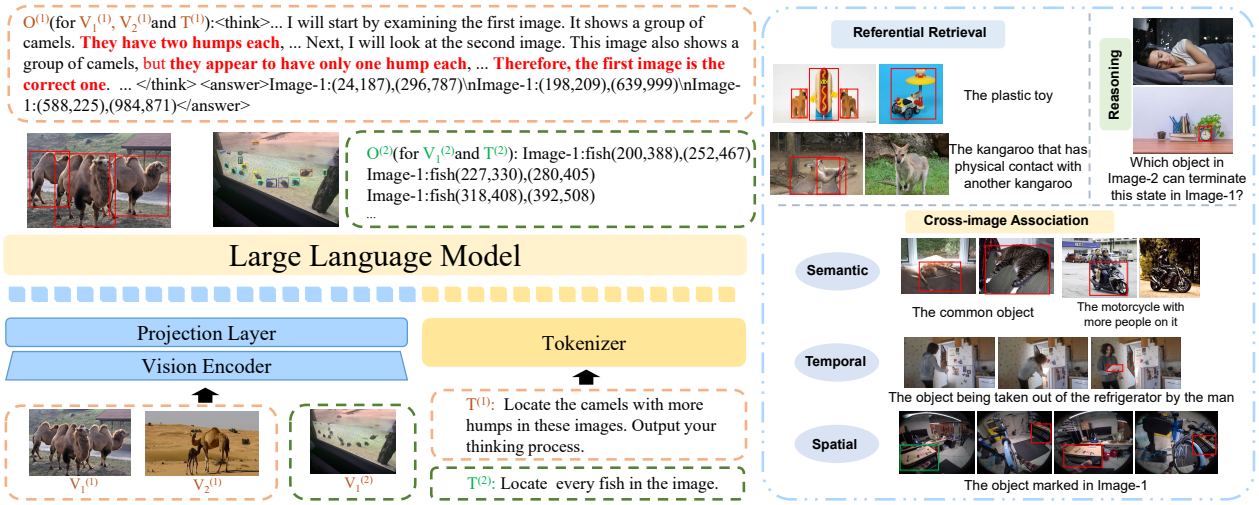
With the emergence of large reasoning models, reinforcement learning has become a research focus for enhancing the reasoning capabilities of LLMs. Recently, DeepSeek-R1 (Guo et al. 2025) achieves a breakthrough in this area by introducing a new rule-based RL algorithm. Inspired by the success of DeepSeek-R1, a series of works apply the rule-based GRPO method to vision-language domain. Among these, some focus on visual reasoning tasks such as mathematical reasoning (Huang et al. 2025; Deng et al. 2025; Wang et al. 2025; Zhang et al. 2025; Yang et al. 2025). Some other studies investigate visual perception tasks including visual grounding (Shen et al. 2025; Zhan et al. 2025; Ma et al. 2025; Liu et al. 2025; Yu et al. 2025), which are closely related to our work. However, they are limited to single-image scenarios, lacking the ability to perform precise grounding across multiple images. In this work, we explore the R1-like paradigm's potential in multi-image grounding.

Methodology

Overview

We begin by defining the generalized multi-image grounding task and outlining GeM-VG. Given an image sequence $V = \{V_1, \ldots, V_m\}$ and a textual instruction T, the model $\mathcal{M}$ is required to localize all relevant instances, outputting bounding boxes with image indices:

$$O = \{(b_1, i_1), \ldots, (b_n, i_n)\} = \mathcal{M}(V, T), \qquad (1)$$



(a) Framework of GeM-VG

(b) Categorization of multi-image grounding tasks

Figure 2: **(Left)** Overview of GeM-VG. GeM-VG is capable of reasoning over multiple visual contexts and localizing all referring instances. **(Right)** We provide a systematic taxonomy of generalized multi-image grounding scenarios.

where b_k is the k-th bounding box and $i_k \in \{1, \ldots, m\}$ denotes its source image.

As shown in Figure 2(a), GeM-VG is built upon Qwen2-VL with three components: a vision encoder, a projection layer, and an LLM. The images V are first encoded by the vision encoder and then projected into the word embedding space to obtain visual tokens H_v, which are concatenated with instruction embeddings H_{ins} and fed into the LLM. A unified output format is adopted to support multi-object localization across multiple images.

We first conduct supervised finetuning to equip the base model with basic multi-image perception and reasoning abilities. Motivated by the success of R1-like algorithm in both perception and reasoning tasks (Deng et al. 2025; Wang et al. 2025; Shen et al. 2025; Zhan et al. 2025; Liu et al. 2025; Ma et al. 2025), we introduce this paradigm to multi-image grounding and design a rule-based reward. Reinforcement learning further improves the grounding capability. The following subsections detail the data construction and reinforcement finetuning strategy.

MG-Data-240K

To mitigate the lack of multi-target grounding capabilities in existing multi-image grounding models, we construct a new dataset, MG-Data-240K, to support broader scenarios. This dataset addresses the limitations of existing datasets in target quantity and image relationships.

Task Taxonomy To cover diverse multi-image grounding scenarios, we categorize tasks based on their cognitive demands and image relationships. As shown in Figure 2(b), tasks are divided into three main types like MC-Bench (Xu, Zhu, and Yang 2024): referential retrieval grounding, where the instructions identify instances using explicit expression without cross-image referring; cross-image association grounding, where the targets are identified through cross-image correspondences; and reasoning grounding, which requires commonsense or external knowledge. Within cross-image association grounding, we further distinguish semantic, temporal, and spatial relations, inspired by cognitive theories on semantic, episodic, and spatial memory (Baddeley 2000; Moscovitch et al. 2006).

Type	Num.	Source
Referring Retrieval Grounding	97K	D^3 (Xie et al. 2023), COCO (Lin et al. 2014)
Semantic Association Grounding	77K	COCO
Spatial Association Grounding	20K	Ego-Exo4D (Grauman et al. 2024), MVTrack (Xu et al. 2025)
Temporal Association Grounding	46K	STAR (Wu et al. 2024)

Table 1: Details of the constructed training dataset.

Data Collection Existing training datasets (Li et al. 2025; Bai et al. 2025b) lack multi-target samples and mainly involve semantic relationships. To address this, we aim to expand the data with more multi-target samples, multi-view images, and diverse practical tasks. Guided by the taxonomy, we select multiple image and video datasets as sources. We form multi-image groups by pairing related images or extracting key frames, depending on task types and annotations. Task instructions are generated using predefined templates combined with metadata such as object labels and QA pairs from the source datasets. Table 1 summarizes the data sources and statistics.

Reinforcement Learning for Generalized Multi-Image Grounding

In this subsection, we detail the reinforcement learning algorithm used to enhance the model's generalized multi-image grounding capability. Our method is an extension of the

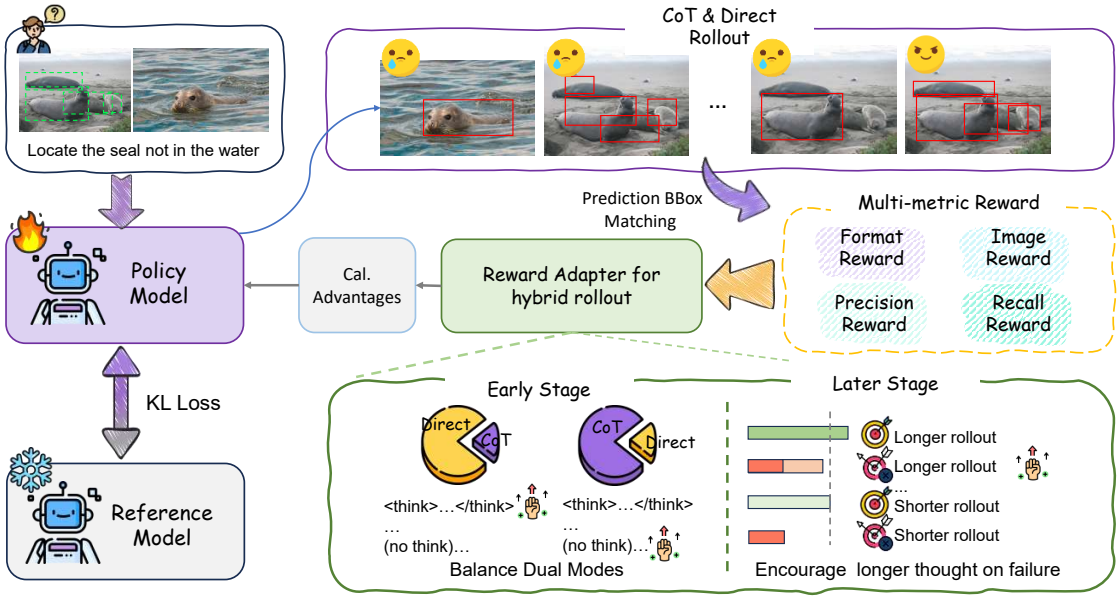


Figure 3: Framework of our reinforcement finetuning. Considering several failure types in multi-image grounding, we design a multi-dimensional rule-based reward. Before computing the advantages, we apply a reward adjustment strategy to facilitate joint optimization between both answering paradigms.

Group Relative Policy Optimization (GRPO) (Shao et al. 2024) algorithm to the visual grounding field.

Given an input question q, the policy model π_θ generates a group of N candidate completions $\{o_1, o_2, ..., o_N\}$. For each completion o_i, a rule-based reward function $R(q, a, o_i)$ computes a scalar reward r_i, where a denotes the ground truth. To assess the relative quality of the completions in the same group, these rewards are used to compute the advantages:

$$A_i = \frac{r_i - mean(\{r_j\}_{j=1}^N)}{std(\{r_j\}_{j=1}^N)} \tag{2}$$

Then the policy model π_θ is updated by maximizing the following objective function:

$$\mathcal{J}_{GRPO}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i - \beta\mathcal{KL}(\pi_\theta(o_i|q)|\pi_{ref}(o_i|q)) \tag{3}$$

Where β is a hyper-parameter to control the KL-divergence.

Reward Function Previous methods typically formulate multi-image grounding as a single-target localization problem and adopt a simple IoU score as accuracy reward. To tackle more generalized scenarios, we design a reward function applicable to arbitrary numbers of instances. As illustrated in Figure 3, we propose a reward function that evaluates the output quality from multiple perspectives, incorporating format reward, image reward, precision reward and recall reward.

- **Format Reward:** The reward R_{format} ensures that each completion o_i follows a required format. Specifically, each prediction bounding box must be listed as: `Image-N:<object_ref_start>description<object_ref_start><box_start>(x1,y1),(x2,y2)<box_end>`. In addition, the corresponding image index must be valid numeric values within bounds. The reward is 1 if the format is satisfied and 0 otherwise.
- **Image Reward:** The reward R_{image} evaluates whether the model correctly identifies which images contain the targets, regardless of the precise location of instances.
- **Precision Reward:** The precision reward assesses the quality of predicted bounding boxes at the instance level. Before computing rewards, we perform bipartite matching between predictions and ground-truth instances. After matching, each predicted bounding box $\hat{b}_i$ is associated with a ground-truth bounding box $b_{\text{match}(i)}$, and an Intersection over Union (IoU) score IoU_i. To encourage the model to generate more precise bounding boxes, the precision reward is defined as the average IoU score over all matched prediction instances:

$$R_{precision} = \frac{1}{M}\sum_{i=1}^{M}\text{IoU}_i \tag{4}$$

- **Recall Reward:** As a complement to the precision reward, the recall reward encourages the model to output all instances of interest. It is defined as the proportion of ground-truth instances successfully matched by a predicted box with an IoU score above a threshold τ (set to 0.5 in our experiments):

$$R_{recall} = \frac{1}{num(GT)}\sum_{i=1}^{M}\mathbb{I}(\text{IoU}_i \geq \tau) \tag{5}$$

The overall reward is computed as:

$$R = R_{format} + R_{image} + R_{precision} + R_{recall} \quad (6)$$

Reward-Modulated Hybrid Finetuning Strategy

Multi-image grounding tasks require models to comprehend complex instructions and visual inputs. The previous approach (Bai et al. 2025b) introduces explicit CoT reasoning processes and significantly improves performance on reasoning grounding tasks. However, CoT-only training can be less effective for tasks that emphasize fine-grained visual perception over complex reasoning. For instance, as illustrated in Figure 1, the region locating task in MIG-Bench(Li et al. 2025) often involves detecting subtle visual cues or distinguishing among highly similar objects, where fine-grained perception and discrimination are critical. In such cases, ambiguous or imprecise descriptions may distract the model's attention away from critical visual information.

To leverage both paradigms, we mix CoT and direct-answer samples during SFT and remove prompts that enforce a fixed answering style. However, we observe that the model quickly converges to direct answering during the subsequent RL training. We ascribe this to the model's reliance on CoT-specific prompts. When such prompts are removed, the model tends to direct prediction. To mitigate this, we propose a reward-modulated hybrid finetuning strategy (Figure 3) consisting of two components: balancing dual modes in the early training stage and encouraging longer thoughts on failure in the later stage.

Balancing Dual Modes (Early Stage) To avoid premature convergence to a single answering mode during training, we introduce a reward modulation mechanism based on the proportion of different modes. Completions are categorized into four types: accurate/inaccurate CoT and accurate/inaccurate direct. A completion is considered accurate if its instance-level average reward exceeds 0.5.

Let p denote the proportion of CoT completions in a batch, and θ be the target balancing ratio. The adjustment magnitudes for accurate and inaccurate completions are δ_1 and δ_2, respectively. A scaling factor α controls the overall adjustment strength. The adjustment of reward is:

$$R_{\text{adjust}}(o_i) = \begin{cases} \alpha \cdot (\theta - p) \cdot \delta_1, & \text{if } \mathbb{I}_{\text{acc-cot}}(o_i) = 1 \wedge p < \theta \\ \alpha \cdot (\theta - p) \cdot \delta_2, & \text{if } \mathbb{I}_{\text{inacc-cot}}(o_i) = 1 \wedge p < \theta \\ \alpha \cdot (p - \theta) \cdot \delta_1, & \text{if } \mathbb{I}_{\text{acc-direct}}(o_i) = 1 \wedge p > \theta \\ \alpha \cdot (p - \theta) \cdot \delta_2, & \text{if } \mathbb{I}_{\text{inacc-direct}}(o_i) = 1 \wedge p > \theta \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

In our experiments, we set $\theta = 0.5$, $\alpha = 2.0$, $\delta_1 = 1.0$, and $\delta_2 = 0.5$. This adjustment strategy encourages a balanced behavior between both answering modes.

Encouraging Longer Thoughts on Failure (Late Stage) The later stage aims to improve the model's performance across both modes. Therefore, the balancing strategy used in the early stage is removed. The optimization relies primarily on the naive reward, incentivizing the generation of responses with higher localization quality. Moreover, we find that the model finetuned using CoT data demonstrates advantages in low-scoring subtasks. Motivated by this, an additional length-aware reward is added to the naive reward for inaccurate completions, encouraging more deliberate reasoning on lower-quality cases. The reward adjustment is computed as:

$$R_{adjust}(o_i) = \begin{cases} \gamma \cdot \tilde{\ell}_i \cdot \Delta_i^{\max}, & \text{if } \mathbb{I}_{\text{inaccurate-cot}}(o_i) = 1 \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

Where $\tilde{\ell}_i = \frac{\ell_i - \min_{j \in G} \ell_j}{\max_{j \in G} \ell_j - \min_{j \in G} \ell_j + \epsilon}$ is the normalized length of the completion o_i within the group G. To ensure that the adjusted rewards of inaccurate completions do not exceed that of accurate ones, the adjustment is capped by $\Delta_i^{\max} = \max(0, r_{\min}^{\text{acc}} - r_i)$, where $r_{\min}^{\text{acc}}$ is the lowest naive reward of accurate completions within the group. The scaling factor γ is set to 0.3 for soft modulation.

During RL optimization, we use the adjusted rewards $R_{adjusted} = R + R_{adjust}$ to calculate advantages.

Experiments

Implementation Details

Training Data The training process comprises three stages. To equip the model with comprehensive multi-image grounding capabilities, we combine the MGrounding-630k dataset with an additional 240k samples constructed from multiple existing datasets to form the training data for Stage1-SFT. To mitigate catastrophic forgetting, we also mix multi-image understanding and single-image grounding data into the training. In Stage2-SFT, we incorporate both CoT and direct-answer annotations derived from UniVG-R1 (Bai et al. 2025b). The RL stage is trained on 26k samples, which are sampled from prior stages.

Benchmarks We conduct comprehensive evaluations across multi-image grounding, single-image grounding, video grounding, and multi-image understanding tasks. For multi-image grounding, we evaluate our model on MIG-Bench and MC-Bench. Among them, MIG-Bench requires localizing a single region in the target image, while MC-Bench involves grounding an unspecified number of instances. For single-image grounding, we evaluate on ODINW (Li et al. 2022), which contains rare categories in real-world scenarios, and LLMSeg (Wang and Ke 2024), a reasoning grounding benchmark. For video grounding, we uniformly sample 6 frames and task the model with localizing within one of these frames, with evaluations conducted on ReasonVOS (Bai et al. 2024) and ReVOS (Yan et al. 2024). To assess the model's general multi-image understanding capabilities, we further incorporate several representative benchmarks including MMIU (Meng et al. 2024), MuirBench (Wang et al. 2024a), MIBench(Liu et al. 2024b) and BLINK (Fu et al. 2024).

Training Configurations We adopt Qwen2-VL-7B (Wang et al. 2024b) as the base model due to its strong multi-image understanding and single-image grounding capabilities. During the SFT stage, we employ a learning rate

Model	Spontaneous Grounding			Referential Grounding							AVG
	Difference		Similarity	Visual Reference				Textual	Visual+Textual		
	Static	Robust	Common	OT	MV	Region	Refer	GG	Reason	Co-Re	
Qwen2-VL-72B (Wang et al. 2024b)	51.13	43.61	73.74	24.54	32.63	19.86	37.37	67.83	50.51	17.94	41.91
Mantis (Jiang et al. 2024)	1.52	0.00	3.31	12.18	2.08	1.00	1.01	10.02	0.00	0.85	3.20
LLaVA-OV-7B (Li et al. 2024a)	6.06	3.19	3.43	0.18	1.04	1.08	9.09	15.43	6.93	0.85	4.73
Minicpm2.6 (Yao et al. 2024)	14.58	2.13	14.34	9.82	2.65	1.75	11.11	20.62	2.97	2.56	7.55
mPLUG-Owl3 (Ye et al. 2024)	18.56	6.38	34.93	8.55	7.64	2.41	7.07	22.85	9.09	5.98	12.35
InternVL2-8B (Team 2024)	8.52	19.15	38.40	19.82	10.07	5.24	34.34	39.79	26.80	7.69	20.98
Qwen2-VL-7B (Wang et al. 2024b)	29.92	36.17	43.07	14.55	9.38	15.54	29.29	63.51	44.33	14.30	33.03
Migician (Li et al. 2025)	70.64	45.74	72.76	67.82	60.07	72.57	75.76	84.12	52.58	33.33	63.54
UniVG-R1 (Bai et al. 2025b)	71.97	**58.51**	**93.13**	76.36	66.32	81.71	82.83	**88.04**	62.89	**44.44**	72.64
GeM-VG	**76.89**	56.38	91.53	**86.55**	**68.75**	**85.70**	**84.85**	86.80	**68.04**	41.03	**74.65**

Table 2: Performance on MIG-Bench. OT, MV, GG and Co-Re respectively means object tracking, multi-view grounding, group grounding and correspondence. The metric is Acc@0.5, which considers a prediction correct if its IoU with the ground truth exceeds 0.5.

Model	AP_{50}^{ref}	AP_{50}^{com}	AP_{50}^{rea}	AP_{50}
Gemini-1.5 Pro (Team et al. 2024)	30.5	30.0	26.1	28.4
CogVLM-G (Wang et al. 2024c)	21.1	19.0	16.5	18.2
Qwen2-VL-7B (Wang et al. 2024b)	22.7	22.4	17.2	20.2
Migician (Li et al. 2025)	20.3	26.4	20.1	23.1
GeM-VG	**42.0**	**33.6**	**28.7**	**32.8**

Table 3: Performance comparison on MC-Bench. The superscripts ref, com and rea denote the results for referring, comparison, reasoning instruction type respectively.

Model	Single image		Video	
	ODINW	LLMSeg	ReasonVOS	ReVOS
Qwen2-VL-7B (Wang et al. 2024b)	32.0	35.53	9.83	23.55
Qwen2.5-VL-7B (Bai et al. 2025a)	37.0	—	—	—
Migician (Li et al. 2025)	21.9	34.68	33.41	39.70
UniVG-R1 (Bai et al. 2025b)	33.3	50.60	58.73	60.03
GeM-VG	**41.1**	**50.90**	**64.41**	**60.52**

Table 4: Single image and video grounding results. For ODINW, we follow the evaluation setting in (Bai et al. 2025a). For video grounding, we follow the setting in (Bai et al. 2025b)

Model	MuirBench	BLINK val	MIBench	MMIU
LLaVA-1.5 (Liu et al. 2024a)	23.46	37.13	26.83	19.20
CogVLM (Wang et al. 2024c)	20.85	41.54	—	23.57
Idefics2-8B (Laurençon et al. 2024)	26.08	—	46.39	27.80
mPLUG-Owl3 (Ye et al. 2024)	39.67	50.30	56.66	21.72
InternVL2-8B (Team 2024)	48.70	50.57	52.91	42.00
Mantis (Jiang et al. 2024)	44.50	49.05	45.09	45.60
LLaVA-OV-7B (Li et al. 2024a)	41.80	48.20	71.29	44.46
MiniCPM-V 2.6 (Yao et al. 2024) 2.6	42.65	51.45	71.09	50.19
Qwen2-VL-7B (Wang et al. 2024b)	39.88	52.35	68.06	54.36
Migician (Li et al. 2025)	57.81	51.53	**71.42**	54.89
UniVG-R1 (Bai et al. 2025b)	44.77	51.55	67.29	53.03
GeM-VG	**58.20**	**52.97**	70.16	**55.01**

Table 5: Multi-image understanding results.

precision	recall	ODINW			
		AerialDrone	Aquarium	EgoHands	thermal
Baseline		2.3	23.7	48.7	40.5
	✓	3.3	19.1	49.3	41.1
✓		2.1	17.3	25.2	**49.3**
✓	✓	**5.3**	**25.6**	**56.6**	45.4

Table 6: Ablation on Reward Design.

of 2e-6 and a total batch size of 128. For RL training, we use the multimodal version of the Open-R1 framework (Chen et al. 2025) with its default configuration. The learning rate is set to 1e-6, and the batch size is 16. For each prompt, 8 completions are sampled, with a maximum completion length of 1024. The training is conducted over 1 epoch on 16 NVIDIA H100(80G) GPUs, consuming approximately 8, 2 and 10 hours for each stage, respectively.

Main Results

Multi-image Grounding To comprehensively evaluate the model's multi-image grounding capability, we conduct experiments on MIG-Bench (Li et al. 2025) and MC-Bench (Xu, Zhu, and Yang 2024). MIG-Bench is a free-form multi-image grounding benchmark that encompasses diverse subtasks, requiring the model to localize a single region of interest in the target image based on visual or textual references. As shown in Table 2, our method achieves the best overall performance. Compared to the previous leading model, GeM-VG shows improvements on subtasks that require discerning perception, such as static difference and region locating, as well as those involving reasoning, such as reasoning grounding. Unlike MIG-Bench, MC-Bench evaluates the model's ability to identify an arbitrary number of instances that match the instruction. The textual instructions in MC-Bench are categorized into three styles: referring, comparison, and reasoning. As shown in Table 3, our model outperforms the base model and Migician by 12.6% and 9.7% respectively, demonstrating superior capability in grounding

Models	Spontaneous Grounding			Referential Grounding							AVG
	Difference		Similarity	Visual Reference				Textual	Visual+Textual		
	Static	Robust	Common	OT	MV	Region	Refer	GG	Reason	Co-Re	
Qwen2-VL-7B (Wang et al. 2024b)	29.92	36.17	43.07	14.55	9.38	15.54	29.29	63.51	44.33	16.24	30.20
SFT (Stage 2)											
CoT	76.52	51.06	89.33	80.18	68.06	80.63	80.81	85.77	**62.89**	**39.32**	71.46
Direct	76.70	50.00	89.20	82.91	69.10	86.03	80.81	87.84	45.36	26.50	69.45
Mix	73.48	52.13	90.67	83.64	68.06	82.46	82.83	87.84	61.86	36.75	71.97
RL (Stage 3)											
w/o mode balancing	77.08	55.32	89.08	86.73	70.41	**86.87**	82.83	87.84	51.55	31.62	71.91
w/o late stage strategy	**79.92**	51.06	**92.39**	**86.91**	69.44	85.95	**85.86**	87.63	52.58	32.48	72.4
only CoT	77.08	54.26	90.92	86.55	**71.53**	85.95	84.85	87.22	59.79	37.61	73.58
early stage+late stage	76.89	**56.38**	91.53	85.82	**71.53**	86.20	**85.86**	**88.66**	**62.89**	37.61	**74.31**

Table 7: Ablation on Hybrid Finetuning Strategy.

multiple instances in multi-image scenarios.

Single-image/Video Grounding We also evaluate our model on single-image and video grounding tasks. As shown in Table 4, our model achieves strong performance across all four benchmarks. Specifically, it outperforms UniVG-R1 by 7.8% on ODINW and 5.68% on ReasonVOS, demonstrating superior multi-object localization, temporal understanding, and reasoning capabilities. More details about the results are provided in the supplements.

Multi-image Understanding Our model not only excels in multi-image grounding, but also achieves competitive performance on general multi-image understanding tasks, as shown in Table 5. While UniVG-R1 significantly improves over previous methods on multi-image grounding, it falls short of Migician in broader multi-image understanding. In contrast, our model consistently achieves superior results across a range of benchmarks. Notably, it outperforms other models on the counting and visual grounding subtasks of MuirBench, demonstrating the effectiveness of our approach for generalized multi-image grounding. More detailed results are demonstrated in the supplements.

Ablation Studies

Effectiveness of Reward Design We conduct ablation studies to investigate the impact of different components. Among them, the format reward mainly ensures expected output format for consistent localization results parsing, and the image reward provides coarse image-level supervision. Therefore, our primary focus is on the effects of the precision and recall rewards. Experiments are conducted on several representative ODINW subsets, encompassing a range of challenging scenarios, such as dense objects, small-scale targets, occlusions, and rare image domains. As shown in Table 6, using only the precision reward leads to significant performance drops on subsets like Aquarium and EgoHands, where missed detections are common due to the lack of incentives to localize all instances. In contrast, removing the precision reward weakens supervision on box quality, resulting in more low-quality predictions and degraded AP in some cases. Combining all reward components consistently improves performance across all subsets, validating the effectiveness of our reward design.

Effectiveness of Hybrid Finetuning Strategy The hybrid finetuning strategy serves as a mechanism to enhance the overall performance of the model across various grounding tasks. During SFT, we compare models trained with pure CoT data, direct-answer data, and their combination on MIG-Bench. As shown in Table 7, CoT-based training improves performance on reasoning-intensive tasks such as Reason and Co-Re, but degrades on perception-oriented tasks like Region. In contrast, combining both paradigms achieves better overall performance, which is adopted as our training recipe.

In the RL stage, we ablate different reward modulation strategies. Without intervention, training quickly collapses to direct-answer mode, leading to limited gains on reasoning tasks. Introducing dual-mode balancing improves performance compared to no intervention by encouraging exploration of reasoning trajectories. Ultimately, by applying a two-stage reward adjustment—balancing both completion modes in early training and shifting toward more thoughts on failure in later stage, the model achieves the best overall performance, outperforming all other variants including the pure CoT mode. These results validate the effectiveness of our hybrid finetuning approach in achieving robust performance across diverse multi-image grounding tasks.

Conclusion

In this paper, we propose GeM-VG, an MLLM capable of generalized multi-image grounding while retaining strong capabilities on single-image grounding and multi-image understanding. To endow and incentivize the base model with multi-image grounding and reasoning capabilities, we introduce the MG-Data-240K dataset and a hybrid finetuning strategy based on R1-like reinforcement learning. Extensive experiments demonstrate that our model achieves superior performance across multi-image grounding, single-image grounding and multi-image understanding benchmarks. We hope this work will encourage further research into developing MLLMs with advanced generalized grounding capabilities to support a wider range of real-world applications.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant No. 2022ZD0160601, the National Natural Science Foundation of China under Grant Nos. 62176254 and 62276260. and the Beijing Natural Science Foundation (No. L247028).

References

Baddeley, A. 2000. The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4(11): 417–423.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025a. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bai, S.; Li, M.; Liu, Y.; Tang, J.; Zhang, H.; Sun, L.; Chu, X.; and Tang, Y. 2025b. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*.

Bai, Z.; He, T.; Mei, H.; Wang, P.; Gao, Z.; Chen, J.; Zhang, Z.; and Shou, M. Z. 2024. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37: 6833–6859.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Chen, L.; Li, L.; Zhao, H.; and Song, Y. 2025. Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than 3.

Deng, Y.; Bansal, H.; Yin, F.; Peng, N.; Wang, W.; and Chang, K.-W. 2025. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*.

Fu, X.; Hu, Y.; Li, B.; Feng, Y.; Wang, H.; Lin, X.; Roth, D.; Smith, N. A.; Ma, W.-C.; and Krishna, R. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 148–166. Springer.

Grauman, K.; Westbury, A.; Torresani, L.; Kitani, K.; Malik, J.; Afouras, T.; Ashutosh, K.; Baiyya, V.; Bansal, S.; Boote, B.; et al. 2024. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19383–19400.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Huang, W.; Jia, B.; Zhai, Z.; Cao, S.; Ye, Z.; Zhao, F.; Xu, Z.; 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*.

Jiang, D.; He, X.; Zeng, H.; Wei, C.; Ku, M.; Liu, Q.; and Chen, W. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Laurençon, H.; Tronchon, L.; Cord, M.; and Sanh, V. 2024. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37: 87874–87907.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10965–10975.

Li, Y.; Huang, H.; Chen, C.; Huang, K.; Huang, C.; Guo, Z.; Liu, Z.; Xu, J.; Li, Y.; Li, R.; et al. 2025. Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. *arXiv preprint arXiv:2501.05767*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26296–26306.

Liu, H.; Zhang, X.; Xu, H.; Shi, Y.; Jiang, C.; Yan, M.; Zhang, J.; Huang, F.; Yuan, C.; Li, B.; et al. 2024b. Mibench: Evaluating multimodal large language models over multiple images. *arXiv preprint arXiv:2407.15272*.

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

Ma, X.; Ding, Z.; Luo, Z.; Chen, C.; Guo, Z.; Wong, D. F.; Feng, X.; and Sun, M. 2025. Deepperception: Advancing r1-like cognitive visual perception in mllms for knowledge-intensive visual grounding. *arXiv preprint arXiv:2503.12797*.

Meng, F.; Wang, J.; Li, C.; Lu, Q.; Tian, H.; Liao, J.; Zhu, X.; Dai, J.; Qiao, Y.; Luo, P.; et al. 2024. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*.

Moscovitch, M.; Nadel, L.; Winocur, G.; Gilboa, A.; and Rosenbaum, R. S. 2006. The cognitive neuroscience of remote episodic, semantic and spatial memory. *Current opinion in neurobiology*, 16(2): 179–190.

Nagaraja, V. K.; Morariu, V. I.; and Davis, L. S. 2016. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, 792–807. Springer.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, 2641–2649.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shen, H.; Liu, P.; Li, J.; Fang, C.; Ma, Y.; Liao, J.; Shen, Q.; Zhang, Z.; Zhao, K.; Zhang, Q.; et al. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Team, O. 2024. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy.

Wang, F.; Fu, X.; Huang, J. Y.; Li, Z.; Liu, Q.; Liu, X.; Ma, M. D.; Xu, N.; Zhou, W.; Zhang, K.; et al. 2024a. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.

Wang, H.; Qu, C.; Huang, Z.; Chu, W.; Lin, F.; and Chen, W. 2025. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*.

Wang, J.; and Ke, L. 2024. Llm-seg: Bridging image segmentation and large language model reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1765–1774.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; XiXuan, S.; et al. 2024c. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37: 121475–121499.

Wu, B.; Yu, S.; Chen, Z.; Tenenbaum, J. B.; and Gan, C. 2024. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*.

Xie, C.; Zhang, Z.; Wu, Y.; Zhu, F.; Zhao, R.; and Liang, S. 2023. Described Object Detection: Liberating Object Detection with Flexible Expressions. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*.

Xu, M.; Zhu, Y.; Jiang, H.; Li, J.; Shen, Z.; Wang, S.; Huang, H.; Wang, X.; Zhang, H.; Yang, Q.; et al. 2025. MITracker: Multi-View Integration for Visual Object Tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 27176–27185.

Xu, Y.; Zhu, L.; and Yang, Y. 2024. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. *arXiv preprint arXiv:2410.12332*.

Yan, C.; Wang, H.; Yan, S.; Jiang, X.; Hu, Y.; Kang, G.; Xie, W.; and Gavves, E. 2024. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, 98–115. Springer.

Yang, Y.; He, X.; Pan, H.; Jiang, X.; Deng, Y.; Yang, X.; Lu, H.; Yin, D.; Rao, F.; Zhu, M.; et al. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

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. *arXiv preprint arXiv:2310.07704*.

Yu, E.; Lin, K.; Zhao, L.; Yin, J.; Wei, Y.; Peng, Y.; Wei, H.; Sun, J.; Han, C.; Ge, Z.; et al. 2025. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *European conference on computer vision*, 69–85. Springer.

Zhan, Y.; Zhu, Y.; Chen, Z.; Yang, F.; Tang, M.; and Wang, J. 2024a. Griffon: Spelling out all object locations at any granularity with large language models. In *European Conference on Computer Vision*, 405–422. Springer.

Zhan, Y.; Zhu, Y.; Zhao, H.; Yang, F.; Tang, M.; and Wang, J. 2024b. Griffon v2: Advancing multimodal perception with high-resolution scaling and visual-language co-referring. *arXiv preprint arXiv:2403.09333*.

Zhan, Y.; Zhu, Y.; Zheng, S.; Zhao, H.; Yang, F.; Tang, M.; and Wang, J. 2025. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *arXiv preprint arXiv:2503.18013*.

Zhang, J.; Huang, J.; Yao, H.; Liu, S.; Zhang, X.; Lu, S.; and Tao, D. 2025. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*.