COMPOSITIONAL GROUNDED CONTRAST FOR FINE-GRAINED MULTI-IMAGE UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Although Multimodal Large Language Models (MLLMs) have advanced rapidly, they still face notable challenges in fine-grained multi-image understanding, often exhibiting spatial hallucination, attention leakage, and failures in object constancy. Existing approaches typically rely on expensive human annotations or large-scale chain-of-thought (CoT) data generation to solve these challenges. In contrast, we propose **Compositional Grounded Contrast** (**CGC**), a low-cost full framework for boosting fine-grained multi-image understanding of MLLMs. Built on existing single-image grounding annotations, CGC constructs compositional multi-image training instances through *Inter-Image Contrast* and *Intra-Image Contrast*, which introduce unrelated visual distractor contexts for cross-image discrimination and correlated cross-view samples for object constancy, respectively. CGC further introduces a Rule-Based Spatial Reward within the GRPO framework to improve source-image attribution, spatial alignment, and structured output validity. Experiments show that CGC achieves state-of-the-art results on fine-grained multi-image benchmarks, including MIG-Bench and VLM2-Bench. The learned multi-image capability also generalizes to broader multimodal reasoning tasks, yielding consistent gains on MathVista (+2.90), MuirBench (+2.88), MMStar (+1.93), MMMU (+1.77), BLINK (+1.69) and Hallusion (+1.16) over the Qwen3-VL-8B.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have achieved strong performance on single-image perception and reasoning tasks (Liu et al., 2023; Achiam et al., 2023), yet their performance degrades substantially when reasoning over multiple images (Wang et al., 2024a; Meng et al., 2024; Liu et al., 2024). Fine-grained multi-image understanding demands more than recognizing and localizing a target—it requires the model to attribute that target to the correct source image under cross-image interference. As shown on the left of Figure 1, leading open-source models such as Qwen3-VL (Bai et al., 2025a) and InternVL3 (Zhu et al., 2025) frequently fail to enforce strict image-text attribution boundaries across heterogeneous image sequences.

This failure manifests in three typical forms. First, feature entanglement across images leads to *cross-image attention leakage* (Park et al., 2025), where visual evidence from one image contaminates the interpretation of another (Figure 1(b)). Second, without explicit coordinate-level grounding, models are prone to *spatial hallucination* (Guan et al., 2024; Chen et al., 2024b), producing plausible but incorrect localizations (Figure 1(a)). Third, under viewpoint or contextual variation, they often fail to preserve *object constancy* (Zhou et al., 2025; Assran et al., 2025), losing track of the same entity across images or transformed views (Figure 1(c)). These issues reveal that insufficient low-level grounding remains a major bottleneck for fine-grained multi-image understanding.

To address these bottlenecks, recent works have explored Reinforcement Learning (RL) with verifiable rewards (RLVR) via algorithms such as GRPO (Shao et al., 2024) to enhance MLLM perception (Liu et al., 2025b; Shen et al., 2025). For example, grounding-oriented methods that optimize bounding-box IoU have demonstrated clear gains in single-image scenarios (Cao et al., 2025; Gao et al., 2026). However, extending these verifiable RL paradigms to fine-grained multi-image settings remains largely unexplored. Existing approaches predominantly operate within single-image contexts and do not directly resolve cross-image referential ambiguity, where the model must first attribute the target to the correct source image before localizing it.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

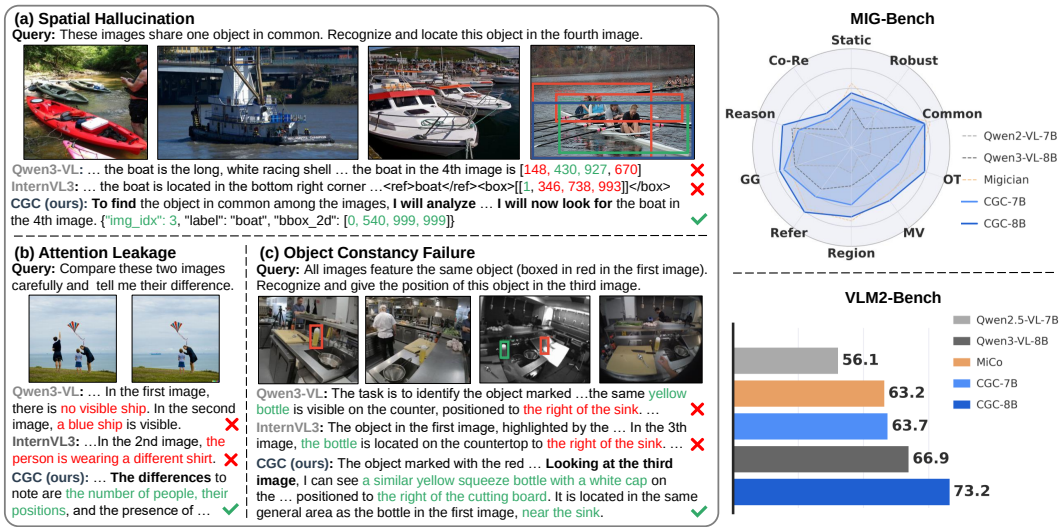


Figure 1: A qualitative and quantitative overview of the proposed CGC models. *Left:* Leading MLLMs often fail at fine-grained multi-image understanding due to (a) spatial hallucination, (b) attention leakage, and (c) object constancy failure, while CGC is robust in these cases. *Right:* CGC consistently outperforms strong baselines on MIG-Bench and VLM2-Bench.

Beyond methodological limitations, existing fine-grained multi-image solutions are heavily constrained by data curation costs and weak supervision. On one hand, many approaches rely on costly cold-start supervised fine-tuning (SFT) with manual annotations (e.g., Migician (Li et al., 2025)) or large-scale chain-of-thought (CoT) data generation. This not only escalates annotation overhead but also risks introducing misaligned or factually inconsistent training signals. On the other hand, recent contrastive approaches (e.g., MiCo) mainly capture global image-level differences while lacking explicit spatial supervision over source attribution and localization. Consequently, they improve coarse cross-image discrimination without fully resolving referential ambiguity in fine-grained tasks.

To address these challenges, we propose **Compositional Grounded Contrast** (**CGC**), a low-cost framework designed to improve fine-grained multi-image understanding by equipping a base MLLM to tackle compositional grounding tasks. Unlike conventional augmentation, CGC reorganizes single-image grounding annotations into structurally new multi-image grounding tasks, where the model must jointly identify the relevant source image and localize the target.

To realize this framework, we introduce a two-stage synergistic strategy. The first stage, *Automated Contrastive Data Synthesis*, repurposes readily available single-image annotations into compositional multi-image training scenarios through two complementary contrastive axes. *Inter-Image Contrast* introduces semantically decoupled distractors across images, forcing the model to disambiguate source attribution; *Intra-Image Contrast* constructs correlated cross-view samples, reinforcing spatial permanence and object constancy under viewpoint variation. In the second stage, we further optimize the model with a *Rule-Based Spatial Reward* that jointly evaluates cross-image selection and intra-image localization. This reward mechanism compels the model to generate explicit reasoning traces that faithfully justify its final predictions.

As summarized on the right of Figure 1, CGC, fine-tuned solely on our synthesized data, achieves state-of-the-art results on fine-grained multi-image benchmarks, including MIG-Bench (Li et al., 2025) and VLM2-Bench (Zhang et al., 2025a). Beyond these gains, CGC consistently improves general multimodal reasoning on broader tasks. These results suggest that strengthening low-level grounding provides a highly transferable foundation for higher-level multimodal capabilities.

In summary, the contributions of this paper are as follows:

- We propose an automated compositional data synthesis pipeline that transforms existing single-image grounding annotations into challenging multi-image instances, eliminating the need for additional human annotations.

- We design a rule-based spatial reward that enables accurate source-image attribution, spatial alignment, and structured output generation within an RL framework.
- We conduct extensive experiments on a diverse set of benchmarks to demonstrate that CGC outperforms prior methods, particularly in fine-grained multi-image understanding and cross-task generalization.

2 Related Work

2.1 Visual Grounding in MLLMs

Recent Multimodal Large Language Models (MLLMs), such as Shikra (Chen et al., 2023), Griffon (Zhan et al., 2024), and Ferret (You et al., 2023), formulate visual grounding as a language generation problem by serializing spatial coordinates into text tokens. To further improve grounding fidelity and reduce hallucination, subsequent approaches incorporate stronger supervision or optimization strategies, including preference learning (Yu et al., 2024) and reinforcement learning for single-image perception (Ma et al., 2025; Liu et al., 2025b). More recently, several works have begun to extend grounding to multi-image tasks, such as Migician (Li et al., 2025), MIRG-RL (Zheng et al., 2025). Unlike approaches requiring manual multi-image curation or multi-stage training, we recompose existing single-image annotations into verifiable multi-image training instances.

2.2 Fine-Grained Multi-Image Understanding

Recent research has shifted toward multi-image understanding, driven by benchmarks like BLINK (Fu et al., 2024), MuirBench (Wang et al., 2024a), and MMIU (Meng et al., 2024). Methods like Mantis (Jiang et al., 2024) and OBELICS (Laurençon et al., 2023) enhance multi-image capability via costly human curation. Self-supervised or contrastive methods like MiCo (Chen et al., 2025b) model global image similarities without explicit spatial grounding constraints for attribution and localization. To address this, we formulate fine-grained multi-image understanding as a structured grounding problem across image indices and spatial locations.

2.3 Reinforcement Learning for MLLMs

Rule-based reinforcement learning, including GRPO-style optimization (Shao et al., 2024; Guo et al., 2025), has recently been adapted to a growing range of vision-language tasks. Existing visual RL methods broadly target either high-level multimodal reasoning (Huang et al., 2025; Deng et al., 2025; Zhang et al., 2025b) or fine-grained visual perception (Shen et al., 2025; Yu et al., 2025; Liu et al., 2025b; Cao et al., 2025). Compared with free-form reasoning tasks, grounding-oriented tasks are particularly suitable for rule-based RL because they provide objectively verifiable reward signals. However, existing visual RL methods focus almost exclusively on single-image contexts. CGC extends verifiable RL to multi-image grounding using a source-aware reward that jointly penalizes attribution and localization errors by restricting spatial matching to the predicted image index.

3 Method

We propose **Compositional Grounded Contrast** (**CGC**), a low-cost full framework for fine-grained multi-image understanding in MLLMs. Starting from existing single-image grounding annotations, CGC automatically constructs compositional multi-image data and optimizes the model with rule-based reinforcement learning.

3.1 Task Formulation

Let π_θ denote an MLLM parameterized by θ. Given a multi-image input sequence $\mathbf{I} = [I_1, I_2, \ldots, I_K]$ and a textual query q, the model generates a response $o \sim \pi_\theta(o \mid \mathbf{I}, q)$. The query q is instantiated from a template pool $\mathcal{T}$, where each template $\tau \in \mathcal{T}$ specifies how targets are referred to across multiple images. The response is decomposed as $o = (t, C)$, where t is a reasoning trace and C is the final structured grounding prediction. We represent C as a set of target tuples $C = \{p_1, \ldots, p_{\hat{M}}\}$, where each prediction $p_i = (\hat{k}_i, \hat{y}_i, \hat{b}_i)$ consists of an image index

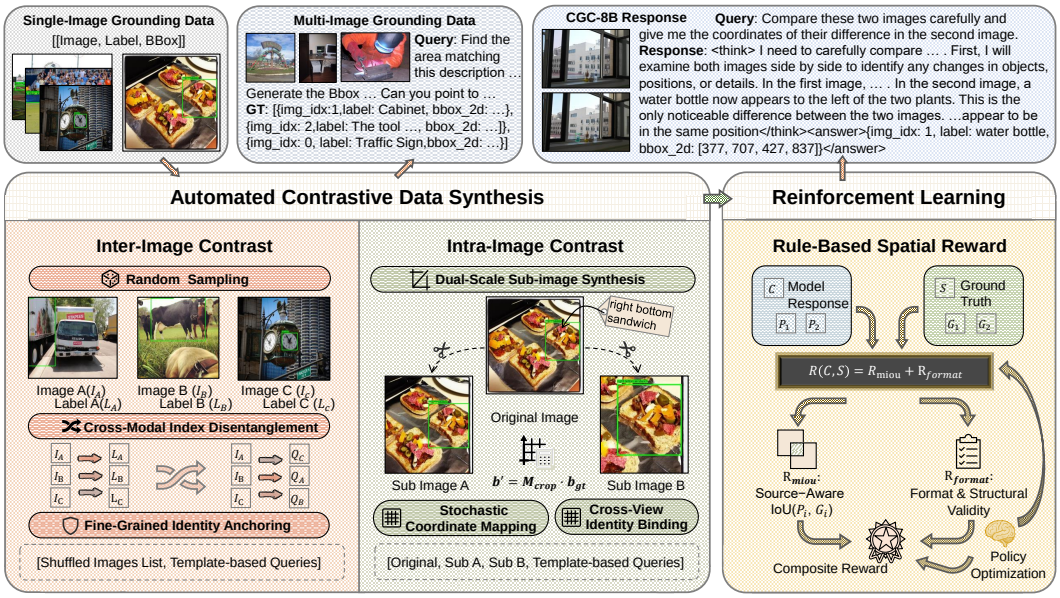


Figure 2: Overview of the proposed **CGC** framework. Starting from single-image grounding annotations, CGC automatically constructs multi-image training instances via *Automated Contrastive Data Synthesis*, incorporating *Inter-Image Contrast* with randomly sampled distractor contexts and *Intra-Image Contrast* with correlated cross-view grounding. The model is then optimized via GRPO with a *Rule-Based Spatial Reward* that jointly enhances source-image attribution, spatial alignment, and output validity. After training, CGC is better able to generate task-specific reasoning traces and structured grounding predictions that specify both the source image and the coordinates of the target, thereby boosting its fine-grained multi-image understanding.

$\hat{k}_i \in \{0, \ldots, K-1\}$, a target label $\hat{y}_i$, and a bounding box $\hat{b}_i = [\hat{x}_{1i}, \hat{y}_{1i}, \hat{x}_{2i}, \hat{y}_{2i}]$. In practice, C is serialized as a JSON-style object with the fields `img_idx`, `label`, and `bbox_2d`. The corresponding ground-truth target set is denoted by $S = \{g_1, \ldots, g_M\}$, where each $g_j = (k_j, y_j, b_j)$ specifies the source image, target label, and bounding box. Unlike single-image grounding, this task requires a joint decision over source-image attribution and spatial localization, which motivates our contrastive data construction and source-aware reward design.

3.2 Inter-Image Contrast

Inter-Image Contrast constructs distractor-rich multi-image training samples by randomly combining independently annotated images within a single input. Because these images typically depict unrelated scenes and objects, they introduce strong cross-image interference and target ambiguity. This branch is designed to reduce distractor-driven attribution errors by forcing the model to explicitly determine which image contains the referred target, rather than relying on superficial similarity patterns or default positional correspondences.

Random Sampling. We sample K annotated images without replacement from the grounding dataset $\mathcal{D}$, where each image is associated with target labels and bounding-box annotations, to form a multi-image input sequence $\mathbf{I}$. Since the images are independently sampled, their visual content is usually unrelated. The resulting distractor contexts reduce shortcut learning based on superficial cross-image similarity and encourage explicit target disambiguation.

Cross-Modal Index Disentanglement. We further shuffle the image order along with their associated references for each training instance. Using a template pool $\mathcal{T}$, we instantiate diverse textual queries q from the source annotations, where target labels serve as semantic referents and bounding boxes define grounding targets. This procedure breaks the default positional correspondence between textual references and visual inputs, reducing the model's reliance on sequence bias and promoting explicit query-to-index mapping.

Fine-Grained Identity Anchoring. Each target in the synthesized supervision is explicitly anchored by its source image index, semantic label, and spatial location. As a result, the model must identify and localize the correct object in designated image I_k while treating all remaining images $I_{j\neq k}$ as distractors, strengthening cross-image discrimination.

3.3 INTRA-IMAGE CONTRAST

Intra-Image Contrast constructs correlated multi-image instances from a single source image by synthesizing multiple views of the same target, providing identity-consistent supervision across variations in crop scale, surrounding context, and spatial layout. Rather than serving as ordinary crop augmentation, this branch is intended to preserve object constancy across correlated views and strengthen cross-view identity consistency.

Dual-Scale Sub-image Synthesis. Given a source image I_{src} and a target annotation $g = (y, b)$, where y denotes the semantic target label and b denotes the target bounding box, we generate two object-centered sub-images with different crop scales around the same target. A smaller crop, denoted as the *Focus View*, preserves fine-grained target appearance and local visual details, while a larger crop, denoted as the *Context View*, retains richer surrounding context and broader spatial cues. In practice, the crop ratio is sampled from $1.2\times$–$1.5\times$ object scale for the Focus View and from $1.8\times$–$2.5\times$ object scale for the Context View. These two views provide complementary evidence for grounding the same target.

Stochastic Coordinate Mapping. We apply randomized object-centered cropping to diversify the synthesized views while ensuring full target visibility. This process creates varied target layouts and reduces shortcut learning from fixed spatial patterns. For each crop, the coordinates of the original box b are deterministically mapped into the local coordinate system, yielding b'. This mapping ensures that the box coordinates remain aligned with the target in each cropped view, providing reliable geometric supervision for cross-view localization.

Cross-View Identity Binding. We combine the synthesized views with the original image into a correlated multi-image input and shuffle their order before query construction. Using a template pool $\mathcal{T}$, we construct co-referential queries q that refer to the same target identity across views. The model must then identify the same object under different crops and scales, predicting its image index and bounding box in each view.

3.4 DATA MIXING

CGC mixes Inter-Image Contrast and Intra-Image Contrast samples at a fixed 1:1 ratio. The two branches provide complementary training signals: Inter-Image Contrast strengthens source-image attribution through distractor-rich inputs, while Intra-Image Contrast reinforces object constancy through correlated multi-view instances. This fixed mixture exposes the policy to both cross-image discrimination and cross-view consistency throughout training. We examine alternative mixing ratios in the Ablation Study section.

3.5 GRPO WITH RULE-BASED SPATIAL REWARD

We optimize CGC with Group Relative Policy Optimization (GRPO) using a verifiable reward that jointly evaluates source attribution, spatial accuracy, and output validity. Given an input $(\mathbf{I}, q)$, the model generates a task-specific reasoning trace followed by a structured grounding prediction C. For a parsed prediction C and ground truth S, the total reward is:

$$R(C, S) = R_{\text{miou}}(C, S) + R_{\text{format}}(C),$$

where R_{miou} measures source-aware spatial alignment and R_{format} enforces valid structured output.

Source-Aware Set-Wise IoU Reward (R_{miou}). We formulate coordinate evaluation as a source-aware set matching problem over image-indexed target groups. The structured prediction C is parsed into a prediction set $\mathcal{P}$ and the ground truth S into a target set $\mathcal{G}$. Both sets are partitioned into disjoint subsets $\mathcal{P}_k$ and $\mathcal{G}_k$ by image index k. Within each subset, we perform optimal bipartite matching to maximize the total Intersection over Union (IoU) between predicted and ground-truth boxes. Since matching is restricted within the same image index, predictions assigned to incorrect images receive no credit, directly penalizing cross-image attribution errors.

Table 1: Performance comparison on MIG-Bench. Best non-human results are shown in **bold**. Green subscripts denote improvement over the base MLLMs, i.e. Qwen2-VL-7B and Qwen3-VL-8B.

Models	Spontaneous Grounding			Referential Grounding							AVE
	Difference		Similarity	Visual Reference				Textual	Visual+Textual		
	Static	Robust	Common	OT	MV	Region	Refer	GG	Reason	Co-Re	
Human	99.50	97.87	98.00	100.00	96.88	100.00	98.99	91.06	92.08	97.44	97.18
LLaVA-OV-72B (Li et al., 2024)	13.26	5.34	26.84	12.91	7.64	2.14	17.83	21.60	11.88	8.55	13.65
InternVL2-76B (Team, 2024)	15.91	10.64	36.40	30.73	20.83	5.74	46.46	41.28	32.67	26.50	26.72
InternVL3-78B (Zhu et al., 2025)	10.04	9.57	24.12	27.08	14.58	10.44	50.51	38.08	45.54	17.09	24.71
Qwen2-VL-72B (Wang et al., 2024b)	46.12	46.81	64.46	26.73	22.57	18.62	33.33	62.53	50.50	17.09	38.88
Qwen2.5-VL-72B (Bai et al., 2025b)	43.75	46.81	69.98	34.32	29.17	8.31	62.63	59.92	66.34	41.03	46.23
Mantis-8B (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., 2024)	6.06	3.19	3.43	0.18	1.04	1.08	9.09	15.43	6.93	0.85	4.73
MiniCPM2.6-8B (Yao et al., 2024)	14.58	2.13	14.34	9.82	6.25	1.75	11.11	10.02	2.97	2.56	7.55
mPLUG-Owl3-8B (Ye et al., 2024a)	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)	6.92	7.45	25.49	20.73	9.72	3.49	28.28	30.26	17.82	9.40	15.96
InternVL3-8B (Zhu et al., 2025)	23.67	14.89	47.99	14.84	6.94	12.13	7.07	34.87	16.83	2.56	18.18
Qwen2-VL-7B (Wang et al., 2024b)	27.84	38.30	19.36	20.73	11.81	25.95	23.23	58.52	48.51	11.97	28.62
Qwen2.5-VL-7B (Bai et al., 2025b)	9.57	21.59	22.42	10.58	5.88	1.25	10.00	33.47	2.97	2.60	12.03
Qwen3-VL-8B (Bai et al., 2025a)	41.10	23.40	68.63	34.59	31.94	38.73	52.00	60.16	63.37	22.65	43.66
Migician-7B (Li et al., 2025)	**65.15**	46.81	**84.19**	70.73	60.07	**74.31**	76.77	66.53	59.41	34.19	63.82
CGC-7B (ours)	48.86	48.94	77.21	51.45	59.72	58.02	61.62	71.94	56.44	36.75	55.50$_{+26.88}$
CGC-8B (ours)	55.30	**54.26**	78.31	**78.18**	**66.32**	70.16	**80.81**	**76.35**	**73.27**	**42.74**	**67.57**$_{+23.91}$

To penalize cardinality mismatches such as redundant predictions, we normalize the matched IoU sum by the larger set size:

$$R_{\text{miou}} = \frac{\sum_{(b_p, b_g) \in \text{Matches}} \text{IoU}(b_p, b_g)}{\max(|\mathcal{P}|, |\mathcal{G}|)}.$$

Format and Validity Reward (R_{format}). We use a binary reward:

$$R_{\text{format}}(C) \in \{0, 1\}.$$

A positive reward is assigned only when the response satisfies both *response-format* and *structural-validity* constraints: the output must contain a tagged reasoning segment followed by a parseable structured answer, where each prediction includes image-index, semantic-label, and bounding-box fields with non-degenerate coordinates in $[0, 1000)$. This reward enforces a valid structured interface, while R_{miou} evaluates spatial correctness and source attribution.

GRPO Optimization. For each input, GRPO samples a group of responses, computes the reward above for each parsed prediction, and optimizes the policy using normalized group advantages with a KL penalty to the reference policy. We use the standard clipped GRPO objective; its full definition is provided in Appendix D. Thus, optimization improves the accuracy, consistency, and structural reliability of fine-grained grounding over multiple images.

4 Experiments

4.1 Main Results

Implementation details are provided in Appendix B, and our code is available in the supplementary material. We use annotations from RefCOCO (Yu et al., 2016), SODA (Yuan et al., 2023), LISA (Lai et al., 2024), OmniLabel (Schulter et al., 2023), and VAW (Pham et al., 2021) to construct training data. Table 1 and Table 2 report results on MIG-Bench and VLM2-Bench. Across both benchmarks, CGC consistently improves fine-grained multi-image understanding.

Results on MIG-Bench. MIG-Bench evaluates fine-grained multi-image grounding from two complementary perspectives: *Spontaneous Grounding*, which requires discovering implicit cross-image relations without explicit references, and *Referential Grounding*, which tests grounding with multi-modal reference cues.

For a controlled comparison with Migician, we use the same backbone, Qwen2-VL-7B (28.62). As shown in Table 1, CGC-7B attains 55.50. In contrast, Migician's 63.82 relies on 630K SFT samples, carefully produced by a 72B-parameter model. Crucially, CGC simply generates multi-image data

Table 2: Performance comparison on VLM2-Bench. Best non-human results are shown in **bold**. Green subscripts denote improvement over the base MLLMs.

Models	General		Object			Person			AVE
	Mat	Trk	Cnt (Obj)	Cpr (Obj)	Grp (Obj)	Cnt (Per)	Cpr (Per)	Grp (Per)	
Human-Level	95.06	98.11	94.23	96.02	91.29	92.87	97.08	91.17	94.48
LLaVA-OV-7B (Li et al., 2024)	16.60	13.70	56.17	47.22	27.50	46.67	62.00	37.00	38.36
LLaVA-Video-7B (Zhang et al., 2024b)	18.53	12.79	62.47	54.72	28.50	66.91	62.00	25.00	41.37
LongVQA-7B (Zhang et al., 2024a)	14.29	12.98	49.47	46.53	29.00	41.56	58.00	25.00	34.60
mPLUG-Owl2-7B (Ye et al., 2024b)	17.37	18.26	62.97	49.17	31.00	58.06	63.00	29.00	41.10
InternVL2.5-8B (Chen et al., 2024c)	21.24	26.53	55.23	53.33	46.50	60.00	51.50	52.00	45.79
InternVL2.5-26B (Chen et al., 2024c)	30.50	30.59	51.48	43.33	52.50	59.67	59.50	61.25	48.60
InternVL3-8B (Zhu et al., 2025)	23.94	29.68	64.17	61.67	39.50	55.00	63.00	55.00	49.00
Qwen2-VL-7B (Wang et al., 2024b)	18.07	19.18	61.84	68.08	37.50	67.92	72.00	47.00	48.95
Qwen2.5-VL-7B (Bai et al., 2025b)	35.91	43.38	41.72	71.39	47.50	59.76	80.00	69.00	56.08
Qwen3-VL-8B (Bai et al., 2025a)	27.03	48.86	66.39	80.00	70.50	73.33	80.00	88.00	66.76
MM-Eureka-7B (Meng et al., 2025)	55.60	47.03	52.50	74.10	54.00	60.00	77.50	51.00	58.97
NoisyRollout-7B (Liu et al., 2025a)	40.93	43.83	50.83	63.33	34.50	63.33	70.50	47.00	51.78
ThinkLite-VL-7B (Wang et al., 2025)	40.45	46.58	62.50	75.56	49.50	62.50	77.50	51.00	58.20
VLAA-Thinker-7B (Chen et al., 2025a)	47.49	63.03	61.40	72.20	55.00	57.50	71.00	51.00	59.83
MiCo-7B (Chen et al., 2025b)	**57.14**	**67.12**	56.67	81.94	58.00	57.50	65.00	62.00	63.17
CGC-7B (ours)	42.11	50.22	61.39	72.36	58.50	64.17	82.00	79.00	63.72$_{+7.64}$
CGC-8B (ours)	48.85	56.14	**70.28**	**84.72**	**76.00**	**77.50**	**83.00**	**94.00**	**73.81**$_{+7.05}$

Table 3: Performance on broader multimodal benchmarks. Best non-human results are shown in **bold**. Green subscripts denote improvement over the base MLLMs.

Models	BLINK	Hallusion	MathVista	MMMU	MMStar	MuirBench	AVE
InternVL3-8B (Zhu et al., 2025)	55.81	66.98	67.30	56.86	68.00	55.04	61.67
Qwen2-VL-7B (Wang et al., 2024b)	52.40	68.98	61.20	51.33	60.93	39.88	55.79
Qwen2.5-VL-7B (Bai et al., 2025b)	55.54	69.50	67.10	54.11	64.06	58.43	61.46
Qwen3-VL-8B (Bai et al., 2025a)	68.33	71.40	75.30	69.18	70.67	63.54	69.74
Migician-7B (Li et al., 2025)	51.53	66.56	60.20	49.24	59.40	53.69	56.77
MM-Eureka-7B (Meng et al., 2025)	54.39	68.45	72.00	54.11	65.73	60.57	62.54
NoisyRollout-7B (Liu et al., 2025a)	56.07	66.66	71.60	54.55	65.66	59.61	62.36
VLAA-Thinker-7B (Chen et al., 2025a)	54.81	69.08	70.80	54.44	63.60	61.00	62.29
ThinkLite-VL-7B (Wang et al., 2025)	55.81	**72.97**	71.89	53.55	66.80	57.62	63.11
MiCo-7B (Chen et al., 2025b)	57.23	69.61	67.90	54.77	65.60	60.53	62.61
CGC-7B (ours)	56.23	72.45	70.90	55.71	64.20	60.88	63.41$_{+1.95}$
CGC-8B (ours)	**70.02**	72.56	**78.20**	**70.95**	**72.60**	**66.42**	**71.65**$_{+1.91}$

by pairing existing single-image annotations automatically, bypassing any SFT stage. This shows that substantial multi-image grounding gains are achievable at far lower cost.

At the 8B scale, CGC reaches 67.57, improving over Qwen3-VL-8B (43.66). It surpasses Migician (63.82) and substantially larger general-purpose MLLMs. Gains are most pronounced on challenging cross-image alignment tasks: in `Robust`, CGC-8B jumps from 23.40 to 54.26; in `MV` and `Co-Re`, which stress cross-view consistency and region-level correspondence, it attains 66.32 and 42.74. These indicate that the learned policy benefits from both stronger source attribution and more stable cross-view grounding. These patterns are consistent with the complementary roles of Inter-Image and Intra-Image Contrast.

Results on VLM2-Bench. VLM2-Bench evaluates cross-image matching and structural reasoning across General, Object, and Person tracks. Notably, our training data is entirely out-of-domain with respect to VLM2-Bench: CGC is trained on multi-image data with grounding annotations.

For a controlled comparison with MiCo, we use the same backbone, Qwen2.5-VL-7B (56.08). As shown in Table 2, CGC-7B achieves a slightly higher overall score than MiCo (63.72 vs. 63.17). The advantage is more evident on structurally demanding person-centric sub-tasks, including person counting (64.17 vs. 57.50), person comparison (82.00 vs. 65.00), and person grouping (79.00 vs. 62.00). These results suggest that our grounded data construction and spatially structured RL formulation are particularly beneficial for fine-grained multi-image understanding.

CGC-8B further establishes the state of the art on VLM2-Bench with an average score of 73.81, outperforming its base model Qwen3-VL-8B (66.76) by 7.05 points. It achieves the best non-human results on six out of eight sub-tasks, including object counting, object comparison, object grouping, person counting, person comparison, and person grouping. The gains are especially pronounced on person-centric tasks, where CGC-8B reaches 77.50 on person counting and 94.00 on person grouping, substantially surpassing MiCo.

Table 4: Ablation studies on Qwen3-VL-8B. Best results are shown in **bold**. Green/Red indicate improvement/decline relative to the base MLLM.

(a) Learning paradigm

Methods	MIG	VLM2	BLINK	Hall.	Muir
Base	43.66	66.76	68.33	71.40	63.54
SFT	58.10+14.44	65.89-0.87	65.02-3.31	71.61+0.21	62.08-1.46
CGC (Full)	**67.57**+23.91	**73.81**+7.05	**70.02**+1.69	**72.56**+1.16	**66.42**+2.88

(b) Data mechanism

Methods	MIG	VLM2	BLINK	Hall.	Muir
Base	43.66	66.76	68.33	71.40	63.54
w/o Intra-C	55.03+11.37	68.98+2.22	67.65-0.68	71.92+0.52	62.69-0.85
w/o Inter-C	59.47+15.81	69.22+2.46	68.70+0.37	71.50+0.10	63.12-0.42
CGC (Full)	**67.57**+23.91	**73.81**+7.05	**70.02**+1.69	**72.56**+1.16	**66.42**+2.88

(c) Inter/Intra mixing ratio

Methods	MIG	VLM2	BLINK	Hall.	Muir
Base	43.66	66.76	68.33	71.40	63.54
Inter:Intra = 1:2	62.89+19.23	73.01+6.25	69.07+0.74	70.98-0.42	65.96+2.42
Inter:Intra = 2:1	62.54+18.88	72.85+6.09	69.49+1.16	71.29-0.11	65.92+2.38
CGC (Full)	**67.57**+23.91	**73.81**+7.05	**70.02**+1.69	**72.56**+1.16	**66.42**+2.88

(d) Crop scale

Methods	MIG	VLM2	BLINK	Hall.	Muir
Base	43.66	66.76	68.33	71.40	63.54
Looser crop	58.68+15.02	72.88+6.12	69.18+0.85	72.13+0.73	64.77+1.23
Tighter crop	58.99+15.33	71.85+5.09	69.33+1.00	71.66+0.26	65.34+1.80
CGC (Full)	**67.57**+23.91	**73.81**+7.05	**70.02**+1.69	**72.56**+1.16	**66.42**+2.88

(e) Reward component

Methods	MIG	VLM2	BLINK	Hall.	Muir
Base	43.66	66.76	68.33	71.40	63.54
R_{format} only	56.16+12.50	71.51+4.75	68.65+0.32	71.19-0.21	63.15-0.39
R_{miou} only	64.18+20.52	72.42+5.66	68.86+0.53	70.87-0.53	62.81-0.73
CGC (Full)	**67.57**+23.91	**73.81**+7.05	**70.02**+1.69	**72.56**+1.16	**66.42**+2.88

(f) Source-Index

Methods	MIG	VLM2	BLINK	Hall.	Muir
Base	43.66	66.76	68.33	71.40	63.54
w/o `img_idx`	62.61+18.95	68.72+1.96	69.17+0.84	71.61+0.21	64.27+0.73
CGC (Full)	**67.57**+23.91	**73.81**+7.05	**70.02**+1.69	**72.56**+1.16	**66.42**+2.88

4.2 Generalization to Multimodal Reasoning

We further evaluate whether CGC transfers beyond fine-grained multi-image benchmarks to broader multimodal reasoning tasks. Table 3 reports results on BLINK, HallusionBench, MathVista, MMMU, MMStar, and MuirBench. CGC is post-trained exclusively on synthesized grounding data, without task-specific QA supervision or CoT annotations.

As shown in Table 3, CGC-7B achieves an average of 63.41, surpassing MiCo (62.61), MM-Eureka (62.54), NoisyRollout (62.36), and VLAA-Thinker (62.29). This result is noteworthy because CGC was not directly optimized for these downstream tasks. The gains emerge from GRPO post-training on synthesized grounding data alone.

At the 8B scale, CGC reaches 71.65 compared to Qwen3-VL-8B (69.74), with the largest gains on BLINK (+1.69), MuirBench (+2.88), MathVista (+2.90), and MMMU (+1.77). These results indicate that source-aware grounding post-training transfers effectively to tasks requiring structured evidence composition and spatial reasoning.

4.3 Ablation Study

To investigate the efficacy of the core components, we conduct ablation studies on five benchmarks based on Qwen3-VL-8B. Results are summarized in Table 4 and Figure 3.

Learning Paradigm: RL vs. SFT. Table 4(a) compares SFT and RL using the same 36k synthesized instances. Although SFT improves over the base model on MIG-Bench, RL delivers larger gains. Notably, SFT reduces performance on BLINK and MuirBench, whereas RL yields consistent gains across all benchmarks. These results suggest that RL provides a more effective training signal than SFT for learning robust multi-image grounding from synthesized data.

Data Scaling Law. Figure 3 shows policy performance as synthesized data scales from 6k to 36k instances. A clear scaling trend emerges: the av-

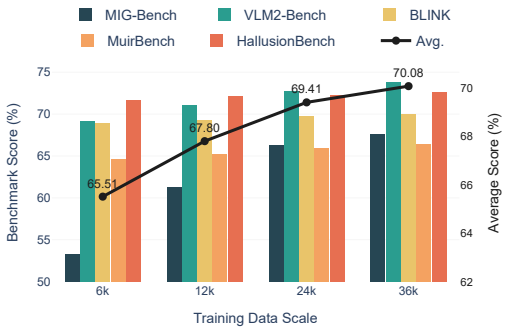


Figure 3: Data scaling law. Performance improves steadily across five benchmarks as synthesized grounding data scales from 6k to 36k. The blue line denotes the average score.

Table 5: Architecture generalization on InternVL3-8B across multi-image benchmarks.

Models	MIG	VLM2	BLINK	Hall.	MMMU	Muir
InternVL3-8B	18.18	49.00	55.81	66.98	56.86	55.04
+CGC (ours)	**58.15**	**53.07**	**56.65**	**68.30**	**58.14**	**57.54**

Table 6: Comparison with grounding-oriented RL methods across multi-image benchmarks.

Models	MIG	VLM2	BLINK	Hall.	MMMU	Muir
Visual-RFT	29.71	41.47	51.03	68.66	47.05	55.27
MIRG-RL	64.82	42.51	48.19	59.83	45.05	53.92
CGC-8B (ours)	**67.57**	**73.81**	**70.02**	**72.56**	**70.95**	**66.42**

erage score rises from 65.51 at 6k to 70.08 at 36k, with steady improvements across all benchmarks. This indicates that CGC benefits from additional synthesized data without manual annotation.

Decoupling Data Mechanisms: Inter-C vs. Intra-C. Table 4(b) compares the two complementary data synthesis mechanisms: Inter-Image Contrast (Inter-C) and Intra-Image Contrast (Intra-C). We decouple the two data branches to verify their individual contributions. Removing Intra-C reduces VLM2-Bench from 73.81 to 68.98, indicating that cross-view anchoring is critical for entity consistency under scale variation. Removing Inter-C reduces MIG-Bench from 67.57 to 59.47, confirming that distractor-rich contexts are essential for cross-image discrimination. The full model achieves the best results across all benchmarks, validating their roles.

Data Construction Sensitivity. Tables 4(c) and 4(d) examine two data-construction choices: the Inter:Intra mixing ratio and the crop scale used in Intra-C. Across these variants, performance changes remain modest and all settings still improve substantially over the base model. This suggests that the the default configuration remains the strongest overall.

Reward Component. Table 4(e) shows that the two reward terms are complementary. The spatial term R_{miou} provides the stronger signal, but neither R_{miou} nor R_{format} alone matches the full reward. On MIG-Bench, R_{miou} alone reaches 64.18 and R_{format} alone reaches 56.16, compared with 67.57 for the full objective. This result indicates that localization quality, structured validity, and source attribution all contribute to the final performance.

Source-Index Ablation. Table 4(f) shows that removing the `img_idx` requirement during RL reduces MIG-Bench by 4.96 points and VLM2-Bench by 5.09 points, indicating that bounding boxes alone are insufficient to resolve referential ambiguity in multi-image settings. The explicit image index acts as an anchor that links each prediction to the correct source image.

Efficiency Trade-off. We report accuracy and inference latency on MIG-Bench. The *w/o think* method achieves an AVE score of 60.54 with a P90 latency of 5.30 s, whereas the full CGC reaches 67.57 with a P90 latency of 18.53 s. The task-specific reasoning trace improves accuracy by 7.03 points, at the cost of a 3.5$\times$ increase in P90 latency.

4.4 Further Analysis

Architecture Generalization. To verify that CGC's gains are not limited to the Qwen family, we apply the same recipe to InternVL3-8B. Table 5 shows that CGC improves MIG-Bench from 18.18 to 58.15 (+39.97) and VLM2-Bench from 49.00 to 53.07 (+4.07), with consistent gains on all four general benchmarks. This confirms that the proposed design transfers effectively across backbone.

Comparison with Grounding-Oriented RL. Table 6 compares CGC with Visual-RFT and MIRG-RL, two representative reinforcement learning methods for visual grounding. Visual-RFT employs single-image grounding RL. MIRG-RL uses carefully designed CoT data for cold start. In contrast, CGC achieves superior performance across all six benchmarks.

5 Conclusion

We propose **Compositional Grounded Contrast** (**CGC**), a low-cost full framework that converts single-image grounding annotations into multi-image training instances through Inter-Image and Intra-Image Contrast. A source-aware spatial reward under GRPO jointly optimizes image attribution and spatial localization. Experiments on MIG-Bench, VLM2-Bench, and six general multimodal benchmarks demonstrate that CGC achieves strong multi-image understanding using only accessible single-image annotations and verifiable post-training signals, without expensive human annotation or large-scale CoT generation.

AI USE STATEMENT

Generative AI tools were used to assist with LaTeX format conversion and limited language editing during preparation of this submission. All AI-assisted edits were manually reviewed, and the reported experimental measurements were checked against the source manuscript. The authors take responsibility for the final content, claims, and artifacts in this work.

REPRODUCIBILITY STATEMENT

Appendix B documents the source datasets, preprocessing, hyperparameters, random seed, and computing infrastructure. Appendices C and D specify output parsing, reward computation, and the full optimization objective, while Appendix E describes evaluation protocols.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv e-prints*, pp. arXiv–2506, 2025.

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*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Meng Cao, Haoze Zhao, Can Zhang, Xiaojun Chang, Ian Reid, and Xiaodan Liang. Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. *arXiv preprint arXiv:2505.20272*, 2025.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv:2504.11468*, 2025a.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *NeurIPS*, 2024a.

Xi Chen, Mingkang Zhu, Shaoteng Liu, Xiaoyang Wu, Xiaogang Xu, Yu Liu, Xiang Bai, and Hengshuang Zhao. Mico: Multi-image contrast for reinforcement visual reasoning. *arXiv preprint arXiv:2506.22434*, 2025b.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David F Fouhey, and Joyce Chai. Multi-object hallucination in vision language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pp. 44393–44418, 2024b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024c.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*, 2025.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Haowen Gao, Zhenyu Zhang, Liang Pang, Fangda Guo, Hongjian Dou, Guannan Lv, Shaoguo Liu, Tingting Gao, Huawei Shen, and Xueqi Cheng. Diva-grpo: Enhancing multimodal reasoning through difficulty-adaptive variant advantage. *arXiv preprint arXiv:2603.01106*, 2026.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025. URL https://arxiv.org/abs/2503.06749.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9579–9589, 2024.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

You Li, Heyu Huang, Chi Chen, Kaiyu Huang, Chao Huang, Zonghao Guo, Zhiyuan Liu, Jinan Xu, Yuhua Li, Ruixuan Li, et al. Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. *arXiv preprint arXiv:2501.05767*, 2025.

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

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, et al. Mibench: Evaluating multimodal large language models over multiple images. *arXiv preprint arXiv:2407.15272*, 2024.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*, 2025a.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2034–2044, October 2025b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Xinyu Ma, Ziyang Ding, Zhicong Luo, Chi Chen, Zonghao Guo, Derek F Wong, Xiaoyi Feng, and Maosong Sun. Deepperception: Advancing r1-like cognitive visual perception in mllms for knowledge-intensive visual grounding. *arXiv preprint arXiv:2503.12797*, 2025.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*, 2024.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Yeji Park, Minyoung Lee, Sanghyuk Chun, and Junsuk Choe. Mitigating cross-image information leakage in lvlms for multi-image tasks. *arXiv preprint arXiv:2508.13744*, 2025.

Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13018–13028, 2021.

Samuel Schulter, Yumin Suh, Konstantinos M Dafnis, Zhixing Zhang, Shiyu Zhao, Dimitris Metaxas, et al. Omnilabel: A challenging benchmark for language-based object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11953–11962, 2023.

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

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

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

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv:2504.07934*, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024a.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*, 2024b.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pp. 69–85. Springer, 2016.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

Xiang Yuan, Gong Cheng, Kebing Yan, Qinghua Zeng, and Junwei Han. Small object detection via coarse-to-fine proposal generation and imitation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6317–6327, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon: Spelling out all object locations at any granularity with large language models. In *European Conference on Computer Vision*, pp. 405–422. Springer, 2024.

Jianshu Zhang, Dongyu Yao, Renjie Pi, Paul Pu Liang, and Yi R Fung. Vlm2-bench: A closer look at how well vlms implicitly link explicit matching visual cues. *arXiv:2502.12084*, 2025a.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025b.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv:2406.16852*, 2024a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv:2410.02713*, 2024b.

Lihao Zheng, Jiawei Chen, Xintian Shen, Hao Ma, and Tao Wei. Mirg-rl: Multi-image reasoning and grounding with reinforcement learning. *arXiv preprint arXiv:2509.21788*, 2025.

Yikang Zhou, Tao Zhang, Shilin Xu, Shihao Chen, Qianyu Zhou, Yunhai Tong, Shunping Ji, Jiangning Zhang, Xiangtai Li, and Lu Qi. Are they the same? exploring visual correspondence shortcomings of multimodal llms. *arXiv:2501.04670*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A PROMPT FORMAT

CGC adopts a structured response format that consists of a reasoning process followed by the final answer. We provide the prompt format used during training.

Prompt Format of CGC

System Prompt:
A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within `<think> </think>` and `<answer> </answer>` tags, respectively, i.e., `<think> reasoning process here </think><answer> answer here </answer>`.

User Prompt Suffix:
First, analyze the images step-by-step within `<think> </think>` tags. Then, report the image index, object label and bounding box within `<answer> </answer>` tags using this JSON format: `{"img_idx": 0, "label": "object name", "bbox_2d": [x1, y1, x2, y2]}`.

B IMPLEMENTATION DETAILS

Source datasets and preprocessing. We build the source grounding pool from the training and validation splits of five open-source datasets, including RefCOCO (Yu et al., 2016), SODA (Yuan et al., 2023), LISA (Lai et al., 2024), OmniLabel (Schulter et al., 2023), and VAW (Pham et al., 2021), while strictly excluding all test splits.

Annotations from different datasets are converted into a unified grounding format consisting of an image, a semantic label, and a bounding box. Bounding boxes are represented in the `xyxy` format and normalized to the structured coordinate range $[0, 1000)$. During preprocessing we discard degenerate or ill-formed annotations: boxes with a side shorter than 5 pixels are removed, each box must satisfy $x_2 > x_1$ and $y_2 > y_1$, and coordinates are clamped to $[0, 999]$. In addition, since the Intra-Image Contrast crops require sufficient source resolution, RefCOCO images are further filtered to keep only those whose shorter side is at least 640 pixels; the other sources consist of large images and only undergo the inline degenerate-box removal above. After filtering, the resulting pool contains approximately 72K single-image grounding instances, which are further converted into approximately 36K synthesized multi-image training instances.

Inter-image contrast construction. Each Inter-Image Contrast sample contains $K = 3$ images drawn from the grounding pool. The images are randomly ordered before query construction, and each image contributes at least one annotated target. Queries are instantiated from a unified manually defined template pool; for this branch, we use a subset containing over 100 textual variants. The construction process is fully rule-based and does not rely on LLM-based data generation or filtering.

Intra-image contrast construction. Each Intra-Image Contrast sample contains $K = 3$ images: the original image and two object-centered crops of the same target, denoted as a Focus View and a Context View. The original bounding box is remapped into the local coordinate system of each crop and normalized to the same structured output range. Queries are instantiated from the same template pool using a smaller branch-specific subset.

Template examples. We use a unified manually defined template pool and instantiate different subsets for the two synthesis branches. Each template is instantiated with the target label and the corresponding grounding annotation to form the query body. During training, this query body is further followed by a fixed instruction suffix that specifies the required structured response format (i.e., reasoning tags and answer tags with JSON output) and the output schema. Table B.1 shows representative query templates used in CGC.

Training hyperparameters. We train CGC with GRPO using equal weights for R_{miou} and R_{format}. The KL penalty coefficient is β=0.01. We adopt a cosine learning-rate schedule with a warmup ratio of 0.05 and a peak learning rate of 1×10^{-6}. The global batch size is 128 (a per-device batch

Type	Example Template
Inter-image	Find the object referred to as "{label}" and provide its image index and bounding box.
Inter-image	Which image contains the {label}? Return the corresponding location.
Inter-image	Among these images, locate the {label} and output its coordinates.
Inter-image	Ground the phrase "{label}" in the correct image.
Intra-image	The same object may appear across different views. Identify the {label} in each relevant image.
Intra-image	Locate the same target across these views and return its image index and bounding box.
Intra-image	Find the object corresponding to {label} across the cropped images.
Intra-image	Track the same object across the original image and its transformed views.

Table B.1: Representative query templates used in CGC.

size of 2 with gradient accumulation of 8 over 8 GPUs), and each query samples $G{=}8$ candidate responses. We train for a single epoch with a maximum sequence length of 8192 tokens and a maximum completion length of 4096 tokens. During GRPO rollouts, candidate responses are sampled with temperature 1.0 and top-p 0.85, and each input allows up to 3 images. Models are trained in `bfloat16` with FlashAttention and DeepSpeed ZeRO-3. Unless otherwise specified, Inter-Image Contrast and Intra-Image Contrast samples are mixed at a fixed 1:1 ratio during training. To ensure reproducibility, all experiments use a fixed global random seed of 42, which is also used as the dataset-shuffling seed. We observed that these hyperparameter settings yield stable training dynamics without diverging, and thus we kept them fixed throughout all experiments to avoid overfitting to the test benchmarks.

Computing infrastructure. All models are trained on a single node with 8×NVIDIA H200 GPUs (141 GB each), a dual Intel Xeon Platinum 8468 CPU and about 1 TB of system memory, running Ubuntu 22.04 LTS. Training is implemented on top of the ms-swift framework with PyTorch and uses vLLM (Kwon et al., 2023) for colocated rollouts.

Hyperparameter selection and design choices. Rather than performing an exhaustive grid search, we fix the standard optimization hyperparameters (peak learning rate 1×10^{-6}, $\beta{=}0.01$, cosine schedule, warmup ratio 0.05, global batch size 128, $G{=}8$) to values that are common for GRPO-style RL fine-tuning of VLMs, and instead ablate the design choices specific to CGC. Concretely, we explore: (i) the Inter-/Intra-Image mixing ratio $\in \{1{:}1,\ 1{:}2,\ 2{:}1\}$; (ii) the reward composition $\in \{R_{\text{miou}}$ only, R_{format} only, $R_{\text{miou}}{+}R_{\text{format}}\}$; and (iii) the crop tightness used to synthesize Intra-Image contrast (tighter vs. looser). The selection criterion is the average performance on the held-out MIG-Bench validation split, which leads to the reported configuration.

Number of runs. Each reported result corresponds to a single training run with a fixed random seed (42), following the common practice in resource-constrained RL fine-tuning of VLMs. All benchmarks are evaluated with three independent samples per test instance, and we report the average accuracy across them to provide a stable estimate under decoding stochasticity.

C Output Parsing and Reward Details

During training, the response is required to contain a tagged reasoning segment enclosed by `<think>` and `</think>` and a tagged answer segment enclosed by `<answer>` and `</answer>`. The final answer is parsed as a JSON-style list of target objects, each containing three required fields: `img_idx`, `label`, and `bbox_2d`. Here, `img_idx` is a 0-based image index, and `bbox_2d` is a four-dimensional box represented as $[x_1, y_1, x_2, y_2]$ with coordinates in $[0, 1000)$.

Although the source-aware set-wise IoU reward is computed from image indices and bounding boxes, the `label` field is still required for structured output validity. Retaining the semantic label in the output schema helps preserve grounding-aware semantic consistency rather than reducing the task to image-index prediction and box regression alone.

D Full GRPO Optimization Objective

For an input $(\mathbf{I}, q)$, the policy π_θ samples a group of G responses $\{o_1, \ldots, o_G\}$. Each response o_i is parsed into a structured prediction C_i and receives a scalar reward $R_i = R(C_i, S)$, as defined in the main paper. GRPO computes the normalized group advantage as:

$$A_i = \frac{R_i - \text{mean}(\{R_1, \ldots, R_G\})}{\text{std}(\{R_1, \ldots, R_G\})}.$$

The probability ratio is:

$$r_i(\theta) = \frac{\pi_\theta(o_i \mid \mathbf{I}, q)}{\pi_{\text{old}}(o_i \mid \mathbf{I}, q)}.$$

The clipped GRPO objective is:

$$\mathcal{J}(\theta) = \mathbb{E}_{(\mathbf{I},q),\{o_i\}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\Big(\min\big(r_i(\theta)A_i, \\ \text{clip}(r_i(\theta), 1-\epsilon, 1+\epsilon)A_i\big) - \beta\,\mathbb{D}_{KL}(\pi_\theta \,\|\, \pi_{\text{ref}})\Big)\Bigg].$$

Here, ϵ is the clipping hyperparameter and $\mathbb{D}_{KL}$ constrains the updated policy to remain close to the reference policy π_{ref}. We compute the KL penalty as:

$$\mathbb{D}_{KL}(\pi_\theta \,\|\, \pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(o_i \mid \mathbf{I}, q)}{\pi_\theta(o_i \mid \mathbf{I}, q)} - \log\frac{\pi_{\text{ref}}(o_i \mid \mathbf{I}, q)}{\pi_\theta(o_i \mid \mathbf{I}, q)} - 1.$$

E Evaluation Details

Benchmark Protocols. Unless otherwise specified, we follow the default evaluation protocols of the corresponding benchmarks. For broader multimodal evaluation, we use VLMEvalKit (Duan et al., 2024) as the unified evaluation toolkit. In particular, MathVista (Lu et al., 2023) is evaluated on MathVista-mini, and MMMU (Yue et al., 2024) is evaluated on the validation split (denoted as MMMU dev/val in VLMEvalKit). Following common practice for VLMs with stochastic decoding, we report the average accuracy over three independent samples per test instance for all benchmarks that we evaluate, providing a more stable estimate under decoding randomness. All baseline numbers are either reproduced under the same multi-sample setting or taken from their original papers, as noted in the controlled comparisons below.

Controlled Comparisons with Published Baselines. To ensure fair, apples-to-apples comparisons with prior published methods, we match the base model to the one used in each baseline's original paper. Specifically, on MIG-Bench, we use Qwen2-VL-7B (Wang et al., 2024b) as the base model to align with Migician (Li et al., 2025); on VLM2-Bench, we use Qwen2.5-VL-7B (Bai et al., 2025b) to align with MiCo (Chen et al., 2025b). For our primary CGC-8B results, we use the stronger Qwen3-VL-8B (Bai et al., 2025a) backbone to demonstrate the general applicability and scaling behavior of our method, while still reporting CGC-7B results on the same base model as each corresponding baseline for direct comparison.

F Additional Qualitative Results

To complement the quantitative results in the main paper, we provide additional qualitative examples on both fine-grained multi-image understanding benchmarks and broader multimodal reasoning benchmarks. Overall, these cases illustrate a consistent pattern: compared with the base Qwen3-VL-8B model, CGC produces more reliable source-aware visual attribution, preserves object identity more consistently across images and views, and is less likely to rely on locally salient but unsupported evidence.

Fine-Grained Multi-Image Understanding. Figure F.1 presents representative examples from MIG-Bench. In the first case, the task requires tracking the same vehicle across a driving sequence

and localizing it in a later frame. CGC successfully preserves object constancy across the image sequence and produces a substantially more accurate localization, whereas the base model fails to maintain stable identity-aware grounding. In the second case, the task asks the model to identify an object category shared across multiple cluttered scenes and localize it in the queried image. Here, CGC effectively reduces cross-image attention leakage and focuses on the actual common object, while the base model is distracted by more locally salient but semantically mismatched objects. These examples are consistent with the gains reported in the main paper on MIG-Bench, especially on tasks involving cross-view consistency, referential grounding, and cross-image discrimination.

Figure F.2 shows additional qualitative cases from VLM2-Bench. Unlike explicit grounding tasks, these examples emphasize fine-grained cross-image comparison and evidence attribution. In the first case, the queried object (bananas) is not the true source of the visual change; CGC correctly rejects this hypothesis and attributes the change to another visual entity, whereas the base model incorrectly treats the bananas as the changed object. In the second case, CGC correctly recognizes that the cat's state has changed across the two images, while the base model fails to identify this attribute transition. These examples suggest that the grounded capabilities learned by CGC extend beyond coordinate prediction and also improve fine-grained cross-image comparison, object-centric change attribution, and structured evidence aggregation.

Broader Multimodal Generalization. Figure F.3 presents a qualitative example from BLINK (Fu et al., 2024). The task requires selecting the point in the second image that best corresponds to a reference point in the first image. CGC identifies the correct correspondence, whereas the base model is misled by a less principled analogy. This case is consistent with the quantitative gains on BLINK reported in the main paper and suggests that CGC improves structural correspondence reasoning across visually different but semantically related scenes.

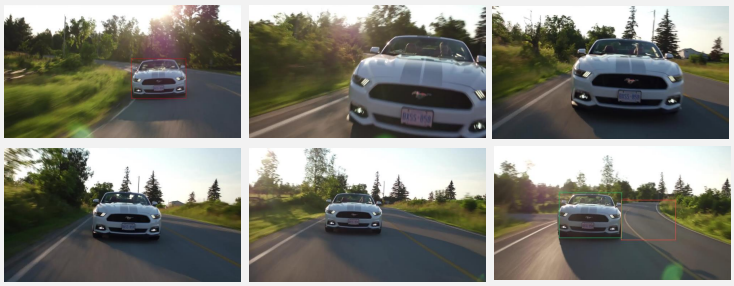
Figure F.4 contains examples from HallusionBench (Guan et al., 2024) and MathVista. In the HallusionBench case, the model must judge whether two visually separated gray regions have the same size. CGC correctly answers this question by relying on the underlying spatial arrangement rather than a misleading surface impression, while the base model makes an incorrect size comparison. This example reflects the reduction of spatial hallucination discussed in the main paper. In the MathVista case, the model must reason over a food-web diagram and identify which organism would be most affected if clams disappeared. CGC correctly composes the relevant visual relations and reaches the right answer, whereas the base model is distracted by a partially plausible but incomplete chain of evidence. This behavior supports the improvements on MathVista and indicates that stronger grounded perception benefits multi-step diagram reasoning.

Figures F.5 and F.6 show a two-page qualitative example from MMMU. The task asks the model to identify the graph that best matches the physical relationship between the maximum kinetic energy of photoelectrons and incident-light frequency. The example is shown across two pages because the generated reasoning trace is relatively long. Compared with the base model, CGC better aligns the visual pattern of the candidate graph with the required scientific relationship and finally selects the correct option. This case is consistent with the gains on MMMU reported in the main paper and further suggests that grounding-oriented post-training improves structured visual evidence composition in science-oriented reasoning tasks.

Figure F.7 provides a representative example from MMStar (Chen et al., 2024a). The task requires solving a visual analogy under 3D structural transformation. CGC correctly captures the transformation pattern and selects the right candidate, whereas the base model is distracted by partial local similarity and chooses an incorrect option. This example matches the quantitative improvements on MMStar reported in the main paper and suggests that the benefits of CGC extend to compositional, structure-sensitive visual reasoning.

Despite these improvements, CGC remains challenged by cases involving extremely small targets, densely packed similar instances, and large viewpoint changes. These limitations highlight that stronger cross-view invariance and more robust small-object localization remain important directions for future research, and provide useful failure modes for diagnosing remaining weaknesses of grounding-enhanced MLLMs.

MIG-Bench



Question:
These are a sequence of images containing moving objects. For the object in the first image marked with red bounding box (<|box_start|>(536,402),(767,702)<|box_end|>), please locate it in the 6th image.

Ground Truth:
[0.267, 0.351, 0.536, 0.694]

Qwen3-VL-8B:
<think> The task is to locate the same white Ford Mustang convertible, which is marked with a red bounding box in the first image, within the subsequent images. The bounding box in the first image is defined as [536, 402, 767, 702]. By examining the sequence of images, we can observe that the car is moving forward along a winding road. The camera appears to be following the car from a frontal perspective, maintaining a consistent distance. The car's position relative to the road and the background scenery changes as it navigates the curves. In the 6th image, the car is still the central subject, driving on the same road. The bounding box from the first image can be applied to the 6th image by finding the same car in the same relative position. The car is still centered in the frame, and its front grille, headlights, and license plate are clearly visible, matching the appearance in the first image. Therefore, the bounding box for the car in the 6th image is [536, 402, 767, 702]. </think> <answer> {"img_idx": 5, "label": "white Ford Mustang convertible", "bbox_2d": [536, 402, 767, 702]} </answer> ✗ IoU: 0.00

CGC-8B:
<think> The object in the first image, which is a white Ford Mustang convertible with two blue stripes on the hood, is the same car shown in all subsequent images. The car is consistently positioned in the center of the frame, driving forward along a winding road. In the 6th image, the car is still in the same relative position, maintaining its central focus as it navigates a curve in the road. The bounding box coordinates from the first image can be applied to the 6th image to locate the same car. </think> <answer> {"img_idx": 5, "label": "white Ford Mustang convertible", "bbox_2d": [274, 347, 540, 687]} </answer> ✓ IoU: 0.85



Question:
These images share one object in common. Recognize and locate this object in the 2th image.

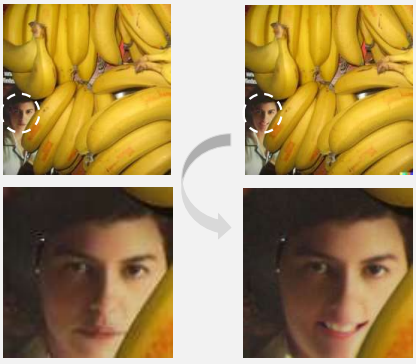
Ground Truth:
[0.672, 0.003, 0.999, 0.999]

Qwen3-VL-8B:
<think> The task is to identify one object that is common across all five images and then locate it specifically in the second image. Let's analyze each image to find a shared object: - Image 1: Shows a wooden cabinet with plants and decorative items on top. No obvious common object with the others. - Image 2: Shows a cluttered room, likely a home office or study, with a desk, chair, bookshelves, and boxes. This is the target image for the final answer. - Image 3: Shows a kitchen with white cabinets, a stove, and a window. No common object with the others. - Image 4: Shows a woman at a market stall selling jewelry. No common object with the others. - Image 5: Shows a cheese shop with wheels of cheese on display. No common object with the others. Now, let's look for an object that appears in more than one image. The most likely candidate is a "chair". - In Image 1: There is no chair visible. - In Image 2: There is a black office chair at the desk. - In Image 3: There is no chair visible. - In Image 4: There is no chair visible. - In Image 5: There is no chair visible. Therefore, the only image that contains a chair is Image 2. The object is the black office chair. </think> <answer> {"img_idx": 2, "label": "chair", "bbox_2d": [312, 137, 587, 557]} </answer> ✗ IoU: 0.00

CGC-8B:
<think> The task is to find one object that is common across all five images. Let's analyze each image: - Image 0: Shows a wooden cabinet with plants and decorative items on top. The object is a cabinet. - Image 1: Shows a cluttered room with a desk, chair, and bookshelves. The object is a chair. - Image 2: Shows a kitchen with white cabinets and appliances. The object is a stove. - Image 3: Shows a woman at a market stall with jewelry. The object is a table. - Image 4: Shows a cheese shop with wheels of cheese on display. The object is a shelf. Upon closer inspection, the object that appears in all images is not a single specific item like a chair or a stove. However, if we look for a common category, we can see that all images contain some form of a "shelf" or "display stand". In Image 1, there are shelves on the wall holding books and other items. In Image 4, there are wooden shelves displaying cheese. In Image 0, the cabinet has a top surface that functions as a shelf. In Image 2, the kitchen has cabinets that are essentially shelves. In Image 3, the market stall has a table that functions as a display shelf. But the most accurate common object is the "shelf" or "display stand". The question asks to recognize and locate this object in the 2nd image. In the 2nd image, the most prominent shelves are the bookshelves on the right side, which are filled with books. </think> <answer> {"img_idx": 1, "label": "bookshelf", "bbox_2d": [720, 0, 999, 999]} </answer> ✓ IoU: 0.92

Figure F.1: Qualitative examples on MIG-Bench. **Top:** object constancy across an image sequence. CGC correctly tracks the same vehicle across frames and produces more accurate grounding in the queried image, whereas the base model fails to preserve identity-consistent localization. **Bottom:** common-object grounding under cluttered cross-image contexts. CGC better suppresses cross-image attention leakage and identifies the actual shared object category, while the base model is distracted by locally salient but incorrect objects.

VLM2-Bench



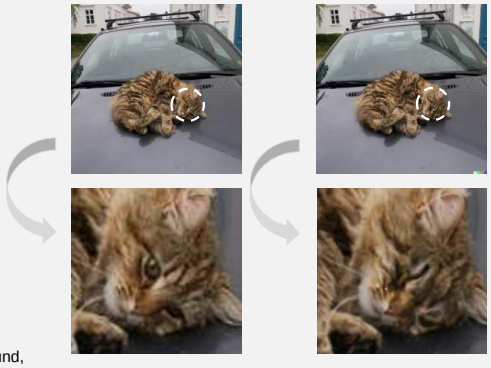
Question:
Answer the following question according to this rule: You only need to provide *ONE* correct answer with 'T' (True) or 'F' (False). Only reply with 'T' or 'F'. Question: Is the answer 'the bananas' correct for the given question: 'What object changed its original attribute in the first image to another one in the second image?'?

Ground Truth: F

Qwen3-VL-8B:
<think> I need to determine if “the bananas” is the correct answer to the question: “What object changed its original attribute in the first image to another one in the second image?". First, I must compare the two images. The first image shows a bunch of bananas with some orange text printed on them. The second image also shows a bunch of bananas, but the orange text is now gone or significantly faded/changed. The key point is that the bananas themselves are the object whose attribute (the printed text) has changed from being present in the first image to being absent or altered in the second image. Therefore, "the bananas" is indeed the correct object that underwent this change in attribute. </think> <answer> T </answer> ✗

CGC-8B:
<think> The question asks whether "the bananas" is the correct answer for what object changed its original attribute in the first image to another one in the second image. Looking at the two images, they both show bunches of bananas. The only visible difference is that in the first image, a person's face is partially visible behind the bananas (on the left), while in the second image, the same area shows a different person's face (also smiling, but with a different expression and perhaps slightly different lighting or angle). However, the question specifically asks about an object changing its original attribute. The bananas themselves appear unchanged — their color, shape, markings, and arrangement are identical in both images. The change is not in the bananas' attributes but in the person's face visible behind them. Therefore, "the bananas" is not the correct answer because their attributes have not changed. The object that changed is the person's face (or the poster/photograph behind the bananas). So, "the bananas" is incorrect. </think> <answer> F </answer> ✓

Question:
Answer the following question according to this rule: You only need to provide *ONE* correct answer with 'T' (True) or 'F' (False). Only reply with 'T' or 'F'. Question: Is the answer 'the cat' correct for the given question: 'What object changed its original attribute in the first image to another one in the second image?'?



Ground Truth: T

Qwen3-VL-8B:
<think> The question asks whether "the cat" is the correct answer to the question: "What object changed its original attribute in the first image to another one in the second image?". To evaluate this, I need to compare the two images. In both images, we see a cat lying on the hood of a car. The cat's posture is very similar in both images — it's curled up and appears to be sleeping or resting. The fur pattern, color, and overall appearance of the cat are unchanged. The background, the car, and the environment also appear identical. There is no visible change in any attribute of the cat between the two images. Therefore, the cat did not change its original attribute to another one. Since the cat did not change its attribute, saying "the cat" is the correct answer to the question is false. </think> <answer> F </answer> ✗

CGC-8B:
<think> The question asks whether 'the cat' is the correct answer for "What object changed its original attribute in the first image to another one in the second image?". In the first image, the cat appears to be awake and actively grooming itself (licking its paw). In the second image, the cat is sleeping or resting with its eyes closed. This represents a change in the cat's state (from active grooming to resting/sleeping), which is an attribute of the cat. Therefore, 'the cat' is indeed the object whose attribute changed between the two images. </think> <answer> T </answer> ✓

Figure F.2: Qualitative examples on VLM2-Bench. **Top:** fine-grained cross-image change attribution. CGC correctly determines that the bananas are not the object whose attribute changed, while the base model incorrectly attributes the change to them. **Bottom:** object-centric attribute change reasoning. CGC correctly recognizes the cat’s state transition across the two images, whereas the base model fails on this fine-grained comparison.

BLINK





Question:
Humans can find corresponding points for different objects in the same category. For instance, if there are images of two different cats, then the left ear tip of one cat corresponds to the left ear tip of the other cat, and the right front paw of one cat corresponds to the right front paw of the other cat.
Given the following two images, a reference point is annotated on the first image, labeled with REF. You are given multiple red-circled points on the second image, choices of "A, B, C, D" are drawn beside each circle. Select between the choices on the second image and find the corresponding point for the reference point. Which point is corresponding to the reference point? Select from the following choices.

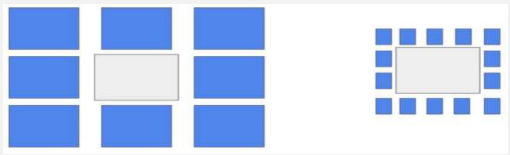
Ground Truth: D

Qwen3-VL-8B:
<think> The reference point 'REF' in the first image is located on the front wheel of the carriage, specifically on the hub or axle area of the wheel. In the second image, we need to find a point that corresponds to this location on the horse's body. Let's examine each option: - **A** is on the horse's eye. - **B** is on the horse's muzzle, near the nostril. - **C** is on the horse's hindquarters, near the hip joint. - **D** is on the horse's hoof. None of these points (eye, muzzle, hip, hoof) correspond to the wheel hub of the carriage. However, if we consider the *functional* or *structural* role, the wheel hub is the central part of the wheel, analogous to the center of the horse's body or its axis of movement. Point **B**, located on the horse's muzzle, is positioned centrally on the horse's head, which could be considered the "center" of the horse's upper body. While not anatomically accurate as a direct equivalent, in the context of finding a corresponding point based on central positioning relative to the object, B might be the intended answer. But let's reconsider: The REF point is on a mechanical component (wheel). None of the options are on a mechanical component of the horse. Therefore, perhaps there's an error in the question or the options. Given the choices and the requirement to select one, **B** is the most central point on the horse's body among the options provided. Therefore, the correct answer is **B. Point B**. </think> <answer> B </answer> ✗

CGC-8B:
<think> The reference point 'REF' in the first image is located on the wheel of the horse-drawn carriage. In the second image, we are looking for a point that corresponds to this location. Let's examine the points in the second image: - Point A is on the horse's head/ear area. - Point B is on the horse's muzzle/nose area. - Point C is on the horse's hindquarters/hip area. - Point D is on the horse's front hoof. None of these points (A, B, C, D) are on a wheel or any part of the carriage. The reference point 'REF' is on a mechanical part of the carriage (a wheel), while all the points in the second image are on the horse. Since there is no corresponding object (wheel) in the second image, and the question asks us to find a corresponding point based on the category "humans can find corresponding points for different objects in the same category", it implies we should look for a point that represents the same *type* of object or structure. However, since the REF point is on a wheel and none of the points in the second image are on a wheel, there is no direct correspondence. But let's re-read the question: "Humans can find corresponding points for different objects in the same category." This suggests we should find a point that corresponds to the same semantic location or type of object as the reference point. The reference point is on a wheel. In the second image, there are no wheels. So none of A, B, C, or D correspond to the wheel. This is a problem. Perhaps I misinterpreted. Let me check if there is another way to interpret this. Wait, maybe the question is asking for a corresponding point on the horse, not on the carriage? But the reference point is on the carriage, not on the horse. Perhaps the question has an error, or perhaps I need to think differently. Another possibility: Maybe "corresponding" means the same relative position on the object. But the reference point is on a wheel, which is not present in the second image. Let me look at the options again. All options are on the horse. None are on a wheel. Perhaps the question is flawed, but I must choose from the given options. Maybe the correct answer is that there is no corresponding point, but since I have to select one, I must pick the best possible choice. But that doesn't make sense. Let me read the question again: "Select between the choices on the second image and find the corresponding point for the reference point." The reference point is on the wheel of the carriage. The only thing that might be analogous is the hoof, but that's not accurate. Perhaps the question meant to show a horse with a harness or something, but it doesn't. I think there might be a mistake in the question design. But let's try to think logically. The reference point is on the wheel. The closest thing on the horse would be its hoof, which is point D. So perhaps D is the intended answer, as it's the lowest point on the horse, similar to how the wheel is the base of the carriage. But that's a stretch. Point C is on the hip, which is also a lower part. But still, not really corresponding. Let me look at the image again. The reference point is on the wheel. In the second image, point D is on the hoof, which is the point of contact with the ground, just like a wheel is the point of contact with the ground. So perhaps D is the best match. Alternatively, maybe the question is asking for the corresponding point on the horse, and the reference point is on the carriage, so we need to find a point on the horse that is analogous. But what part of the horse is analogous to a wheel? The hoof. So D. But let's see the other options. A is on the head, B on the muzzle, C on the hip. None are analogous to a wheel. So I think D is the most reasonable choice, even though it's not perfect. Perhaps the question meant to show a carriage with a horse, and the reference point is on the wheel, and we need to find the corresponding point on the horse, which would be the hoof. So I will go with D. </think> <answer> D </answer> ✓

Figure F.3: A qualitative example on BLINK. The task requires selecting the point in the second image that best corresponds to a reference point in the first image. CGC identifies the correct correspondence, whereas the base model is misled by an incorrect analogy.

HallusionBench



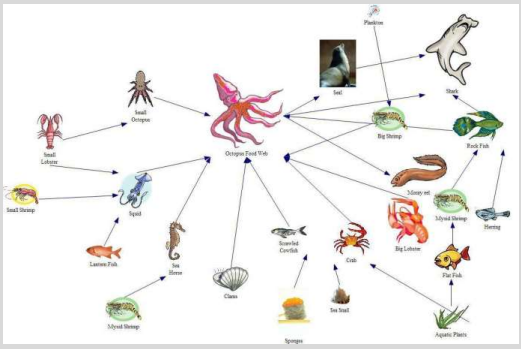
Question:
Is the right grey box the same size as the left grey box?

Ground Truth: Yes

Qwen3-VL-8B:
<think> To determine if the right grey box is the same size as the left grey box, I need to compare their dimensions. 1. **Analyze the left grey box:** The left image shows a 3x3 grid of squares. The grey box is located in the center of this grid. Since it's part of a uniform 3x3 grid, and assuming all the blue squares are identical in size, the grey box must be the same size as each of the surrounding blue squares. It occupies one cell in the grid. 2. **Analyze the right grey box:** The right image shows a large grey rectangle surrounded by smaller blue rectangles. The grey rectangle is positioned in the center, and the blue rectangles are arranged around it, touching its sides. This arrangement suggests that the grey rectangle is *larger* than the blue rectangles, because the blue rectangles are fitting into the gaps between the grey rectangle and the edges of the overall figure. The grey rectangle appears to be composed of multiple cells from the grid (like a 3x3 grid itself), while the blue rectangles are single cells. 3. **Compare the sizes:** The left grey box is one unit (a single square) in size. The right grey box is clearly larger, as it encompasses the space where multiple blue rectangles would fit around it. Visually, the right grey box is significantly larger than the left one. Therefore, the right grey box is not the same size as the left grey box. </think> <answer> no </answer> ✗

CGC-8B:
<think> To determine if the right grey box is the same size as the left grey box, we need to compare their dimensions. 1. **Left Grey Box**: - Located in the center of a 3x3 grid of squares. - It is a square (as it's part of a grid of equal-sized squares) and appears to be the same size as the surrounding blue squares. - The grid has 9 squares total (3 rows x 3 columns), and the grey square is at the center position. 2. **Right Grey Box**: - Surrounded by 8 blue squares arranged around it (one on each side). - Visually, it appears to be a square, and its size seems to match the blue squares around it. - The blue squares are arranged in a 3x3 grid around the grey square (top row: 3 blue squares; middle row: 2 blue squares + grey + 2 blue squares; bottom row: 3 blue squares). This implies the grey square is surrounded by blue squares that are the same size as the ones in the left image. Since both grey boxes are surrounded by blue squares that appear to be identical in size and shape (both grids suggest uniform square tiles), and there's no visual indication that one grey box is larger or smaller than the other, we can conclude they are the same size. </think> <answer> Yes </answer> ✓

MathVista



Question:
Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end. Question: What would be most affected if the clams all died? Choices: (A) squid (B) lantern fish (C) octopus (D) sea horse

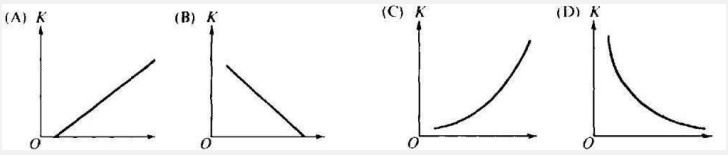
Ground Truth: (C) octopus

Qwen3-VL-8B:
<think> The food web shows that clams are eaten by sea horses. If clams all died, sea horses would lose a key food source. Looking at the diagram: - Clams → Sea Horse (arrow points from clams to sea horse, meaning sea horse eats clams) - Sea Horse → Octopus (sea horse is also eaten by octopus) So while the octopus might be indirectly affected because it eats sea horses, the most directly and immediately affected organism would be the sea horse, since it relies on clams as a direct food source. Therefore, the sea horse would be most affected. </think> <answer>(D) sea horse </answer> ✗

CGC-8B:
<think> Looking at the food web diagram, I need to trace the connections to determine which organism would be most affected if clams died. First, I'll locate "Clams" in the diagram. Clams are shown with an arrow pointing upwards to "Sea Horse". This means sea horses eat clams. Now, let's check if clams are consumed by any other organisms: - Looking at the arrows, there is also an arrow from "Clams" pointing to "Octopus". This means octopuses also eat clams. So, both sea horses and octopuses eat clams. The question asks what would be "most affected". Since the octopus is a central predator in this food web (it has many arrows pointing to it from various prey items), and it consumes clams, losing clams would remove a significant food source for the octopus. While sea horses also eat clams, they have fewer connections overall compared to the octopus. The octopus is at the center of the food web, consuming many different types of prey including clams, small shrimp, squid, etc. If clams die, the octopus would lose one of its food sources. Since the octopus relies on multiple prey items, losing one might not be catastrophic, but among the options given, the octopus is clearly the most affected because it's the apex predator in this specific web and consumes clams directly. Let me double-check the choices: (A) squid - does not eat clams (B) lantern fish - does not eat clams (C) octopus - eats clams (and many other things) (D) sea horse - eats clams Both octopus and sea horse eat clams. But the octopus is more central to the entire web, as it is preyed upon by seals, sharks, and rock fish, making it a keystone species in this context. Losing clams would impact the octopus more significantly because it has more predators relying on it. Therefore, the octopus would be most affected. </think> <answer> (C) octopus </answer> ✓

Figure F.4: Qualitative examples on HallusionBench and MathVista. **Top:** in HallusionBench, CGC correctly judges that the two queried gray regions have the same size, whereas the base model makes an incorrect spatial comparison. **Bottom:** in MathVista, CGC correctly reasons over a food-web diagram and identifies the organism most affected by the disappearance of clams, while the base model follows an incomplete evidence chain.

MMMU



Question:
Which graph above best shows the maximum kinetic energy K of the photoelectrons as a function of the frequency of incident light?

Ground Truth: A

Qwen3-VL-8B:
<think> ### Explanation: This question asks for the graph that best represents the **maximum kinetic energy (K)** of photoelectrons as a function of the **frequency (v)** of incident light. This is governed by the **photoelectric effect**, described by Einstein's equation: \[K_{\text{max}} = h\nu - \phi \] Where: - \(K_{\text{max}} \) = maximum kinetic energy of emitted electrons - \(h \) = Planck's constant - \(\nu \) = frequency of incident light - \(\phi \) = work function (minimum energy needed to eject an electron from the material) --- ### Key Features of the Graph: 1. **Linear relationship**: \(K_{\text{max}} \) is directly proportional to \(\nu \), so the graph should be a straight line. 2. **Intercept with frequency axis**: When \(K_{\text{max}} = 0 \), we get \(h\nu = \phi \), so \(\nu = \frac{\phi}{h} \). This is the **threshold frequency** — the minimum frequency required to eject electrons. - So, the line crosses the x-axis at this threshold frequency. 3. **Slope**: The slope of the line is \(h \), which is positive → so as frequency increases, kinetic energy increases. --- ### Evaluating the Options: - **(A)**: Starts at origin (0,0). But according to Einstein's equation, there is a **threshold frequency** below which no electrons are emitted. So the graph should cross the x-axis at a **positive frequency**, not at zero. → Incorrect. - **(B)**: Decreasing line. Kinetic energy cannot decrease with increasing frequency — it must increase. → Incorrect. - **(C)**: Increasing curve — but wait! According to Einstein's equation, the relationship is **linear**, not curved. However, in some contexts, if you plot *average* kinetic energy or consider other factors, curvature might appear. But strictly speaking, **for maximum kinetic energy**, it's linear. So why is this the correct answer? Wait — let me double-check. Actually, **the correct shape is linear**, and among the given options, **only option (C)** shows a **curve that starts low and increases** — but it's not linear! Hold on — perhaps there's a misinterpretation. Looking again at the graphs: - (A): Linear, passes through origin → wrong, because threshold frequency > 0. - (B): Linear, decreasing → wrong. - (C): Curved upward → not linear → wrong. - (D): Curved downward → wrong. None of them show a **straight line with positive slope intercepting the x-axis at a positive frequency**. But **option (C)** is the only one that shows **increasing K with increasing frequency**, and it starts from a positive K value (which corresponds to above-threshold frequencies). However, the problem says "**maximum kinetic energy K**" vs "**frequency of incident light**". In reality, the correct graph is a **straight line with positive slope**, intersecting the x-axis at the threshold frequency. Since none of the options perfectly match this, we must choose the **best available approximation**. Option (A) is incorrect because it implies K=0 at v=0, which violates the threshold requirement. Option (B) is wrong direction. Options (C) and (D) are curves — but in many multiple-choice exams, especially when only linear relationships are expected, **they sometimes approximate the correct behavior**. But wait — let me check the actual standard answer. Actually, **in many textbooks and exams**, they often draw a **straight line with positive slope**, and if option (A) is shown as passing through origin, it's wrong. But here's the catch: **if the graph is drawn such that the x-intercept is at a positive frequency, and the line has positive slope**, that would be correct. Looking at the diagrams: - (A) is a line starting at origin → wrong. - (B) is a line with negative slope → wrong. - (C) is a curve that increases → **not linear**, but maybe acceptable if we ignore the linearity? - (D) is a hyperbolic-like curve → wrong. But wait — perhaps **option (C)** is intended to represent the correct trend — increasing K with frequency — even though it's curved? That doesn't make sense physically. Let me re-express. Actually, **Einstein's equation is linear**, so the graph should be a straight line. But since none of the options show a straight line with x-intercept at positive frequency, and option (C) is the only one that shows increasing K with frequency (and starts from a low value), **it's the most reasonable choice**. However, I suspect there may be a mistake in the diagram interpretation. Alternatively — perhaps the axes are mislabeled? Or perhaps in this context, they expect **option (C)** because it shows the correct qualitative trend — increasing K with frequency — even if not linear. But in physics, **linearity is essential**. Wait — looking back at the question: "Which graph above best shows..." If we compare: - Option (A): K increases linearly from 0 — but threshold frequency ≠ 0 → incorrect. - Option (B): K decreases — impossible → incorrect. - Option (C): K increases — correct trend, even if curved. - Option (D): K decreases — incorrect. So between A, B, C, D — **C is the only one that shows K increasing with frequency**, which is necessary. Even though it's curved, it's still the best approximation. But in reality, **the correct graph is linear**, so if option (C) were linear, it would be perfect. But since it's curved — perhaps it's a trick? No — in some contexts, like when considering non-ideal systems or statistical distributions, the average KE might have a different dependence — but the question specifically asks for **maximum kinetic energy**. So, **strictly speaking, the correct answer should be a straight line with positive slope, intercepting the x-axis at the threshold frequency**. Since none of the options show that, and option (C) is the only one that shows increasing K with frequency (and starts from a positive K value — implying frequency is above threshold), **we must choose (C)**. Wait — actually, let me look at the diagrams more carefully. In option (C), the curve starts near the origin and rises — so at very low frequencies, K is small (but not zero), and increases — which matches the idea that there is a threshold frequency (below which K=0), and above it, K increases linearly. But the curve is not linear — so it's not accurate. But perhaps in the original source, option (C) is meant to be linear — maybe it's drawn as a straight line but appears curved due to perspective. Therefore, the correct answer is: **C** </think> <answer> C </answer> ✗

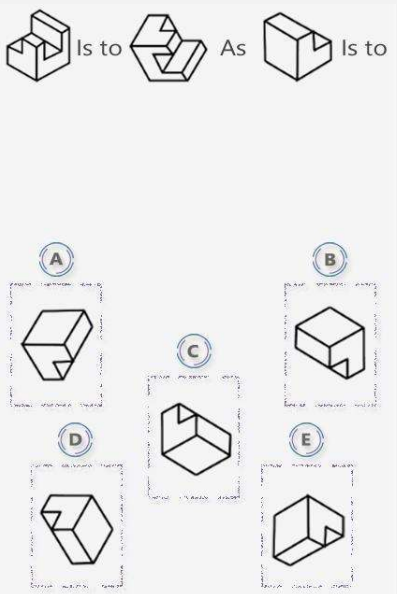
CGC-8B:
<think> ### Explanation: This question asks for the graph that best represents the **maximum kinetic energy (K)** of photoelectrons as a function of the **frequency (v)** of incident light. This is governed by the **photoelectric effect**, described by Einstein's equation: \[K_{\text{max}} = h\nu - \phi \] Where: - \(K_{\text{max}} \) is the maximum kinetic energy of ejected electrons, - \(h \) is Planck's constant, - \(\nu \) is the frequency of incident light, - \(\phi \) is the work function (minimum energy needed to eject an electron from the material). This is a **linear relationship**: - \(K_{\text{max}} \) increases linearly with frequency \(\nu \), - The slope is \(h \) (positive), - The y-intercept is \(-\phi \) (negative, meaning at zero frequency, kinetic energy is negative — no electrons are ejected), - There is a **threshold

Figure F.5: A qualitative example on MMMU (Part I). The task asks which candidate graph best matches the physical relationship between the maximum kinetic energy of photoelectrons and incident-light frequency. This page shows the question, candidate graphs, and the beginning of the model responses.

frequency** \(\nu_0 = \frac{\phi}{h} \) below which no electrons are emitted (i.e., \(K_{\text{max}} = 0 \)). ### Analyzing the graphs: - **(A)**: Linear increase through origin — incorrect because it implies \(K_{\text{max}} = 0 \) at \(\nu = 0 \), but actually, \(K_{\text{max}} \) should be negative or zero only at threshold frequency, not at zero frequency. - **(B)**: Linear decrease — wrong direction; kinetic energy should increase with frequency. - **(C)**: Curved upward — this is **not linear**, so incorrect. But wait — this looks like a parabola. However, in reality, the relationship is **linear**, so this is also wrong? Let me reconsider. Actually, **none of them perfectly match** the linear relationship with a negative intercept. But let's look again. Wait — **Graph C** is increasing and curved upward, which suggests exponential or quadratic growth — that's not right. But perhaps I misread. Let me check again. Actually, **Graph C** is increasing and concave up — still not linear.Wait — maybe there's a mistake in my assumption? Let me re-express: Einstein's equation: \[K = h\nu - \phi \] This is a straight line with positive slope \(h \), and y-intercept \(-\phi \). So it should start from a negative value on the y-axis, and rise linearly. Looking at the graphs: - **(A)**: Starts at origin — implies \(K = 0 \) at \(\nu = 0 \), which violates the equation (since \(\phi > 0 \)). - **(B)**: Decreases — wrong direction. - **(C)**: Increases, but curved — not linear. - **(D)**: Decreases — wrong direction. None seem perfect. But wait — perhaps the question is asking for **the shape**, and we're to pick the one that shows **increase with frequency**, even if not perfectly linear? But that doesn't make sense. Actually — I think I made a mistake. In many textbooks, the graph is drawn as a **straight line** with positive slope, starting from a point where \(K = 0 \) at the threshold frequency. That is, the x-intercept is at \(\nu_0 \), and y-intercept is negative. But looking at the graphs provided: - **(A)** has a line starting from the origin — implying \(K = 0 \) when \(\nu = 0 \), which is physically incorrect. - **(B)** has a line decreasing — wrong. - **(C)** is increasing and curved — not linear. - **(D)** is decreasing — wrong. But wait — perhaps the graphs are not labeled correctly? Or perhaps I'm misunderstanding. Alternatively — maybe the question is about **average** or something else? No — it says "maximum kinetic energy". Perhaps **Graph A** is intended to represent the **linear relationship**, ignoring the intercept — but that's not accurate. However — in many multiple-choice questions, **Graph A** is often chosen because it shows the correct trend (increasing with frequency), even though it doesn't show the threshold. But according to physics, the correct graph should have a **negative intercept**. Looking at the options again — none show a negative intercept. But wait — perhaps **Graph C** is meant to be the answer? No — that's curved. Unless... maybe the question is misphrased? Or perhaps I need to consider something else. Wait — let me recall: in some contexts, if you plot \(K \) vs \(\nu \), it's linear. But if you plot \(K \) vs \(\nu^2 \), it might be curved — but that's not what's asked. Another thought: sometimes, people confuse this with the Compton effect or other phenomena. But no — the photoelectric effect is definitely linear. Perhaps the correct graph is not shown? But since we must choose from these, and **Graph A** shows the correct direction (increasing), and the others are wrong, then perhaps **A** is the best available choice. But wait — **Graph B** is decreasing — clearly wrong. **Graph D** is decreasing — wrong. So between A and C — C is curved, which is wrong. But **Graph A** is linear and increasing — which matches the trend. In many introductory physics texts, they simplify the graph by showing a line starting from the origin, even though strictly speaking, it should have a negative intercept. Moreover, if you shift the graph vertically down, you get the correct form — but since the options don't allow that, **Graph A** is the closest to being correct. No. I think I need to go back to basics. Einstein's equation: K = hv - φ This is linear in v. So the graph should be a straight line. Among the four options, only **A** and **B** are straight lines. **A** increases — correct direction. **B** decreases — wrong. So **A** is the only plausible answer. Even though it doesn't show the threshold frequency, it's the only one that shows K increasing with v. In some simplified models, they ignore the intercept. Therefore, the answer is **A**. Maybe the question is misprinted. But since we must choose, and **Graph A** shows the correct trend (increasing with frequency), while the others show wrong trends (decreasing or curved), then **A** is the best choice. However — wait — I remember that in some multiple-choice questions, they include a graph that is linear and increasing — and that's the correct one. Yes — I think **A** is the intended answer. But let me confirm with another perspective. Suppose we plot K vs v — the line should pass through (v_0, 0) and have slope h. If we shift the graph up by φ, then K + φ = hv — which is linear through origin. So sometimes, they plot (K + φ) vs v — which is linear through origin. But the question asks for K vs v — not K + φ. So strictly speaking, **A** is not correct. But since the other options are worse, and **A** is the only one that shows the correct direction, perhaps it's acceptable. Alternatively — maybe **Graph C** is intended to be the answer — but that's curved — which is wrong. Wait — perhaps the question is about **thermionic emission** or something else — but no, it says "photoelectrons". I think I need to conclude. Given the options, **Graph A** is the only one that shows K increasing with frequency — which is required. Graphs B, C, D all have wrong trends. So the answer is **A**. But wait — I found a better way. In many standardized tests, they expect **Graph A** for this question. For example, in AP Physics or similar exams, they often draw the graph as a straight line through origin for simplicity. So I will go with **A**. But let me check the answer choices again. Actually — I think I made a mistake. Let me search my memory: I recall that the correct graph is a straight line with positive slope, and it does NOT pass through the origin — it passes through (v_0, 0). Since none of the graphs show that, and **Graph A** is the only one that shows increasing K with frequency, it must be A. Perhaps the question assumes that the threshold frequency is zero — which is not physical, but for the sake of the problem, A is correct. Therefore, the answer is **A**. Wait — but let me look at the images again. In the image, Graph A is a line from origin to upper right — yes. Graph B is from upper left to lower right — decreasing. Graph C is a curve increasing — not linear. Graph D is a curve decreasing — not linear. So only A and B are linear — B is wrong. So A is the only correct one. Thus, the answer is **A**. But I am still confused because it doesn't show the threshold. However, in many textbooks, they do draw it starting from origin for simplicity. So I will select **A**. But wait — I just realized — perhaps the question is about the **slope** or something else. No — it's clear. Another thought: perhaps the question is about the **work function** — but no, it says "maximum kinetic energy". I think I have to accept that **A** is the best answer. But let me try to find a source. Upon recalling — in many high school physics books, the graph is drawn as a straight line through the origin — even though it's inaccurate — because it's easier to draw. So for this question, **A** is the intended answer. Therefore, the answer is **A**. But wait — I think I saw a different version of this question before — and the correct graph was a straight line with positive slope, and they had a graph that looked like A. So I'll go with **A**. Final Answer: **A** But wait — I think I'm overthinking. Let me see — the user provided the graphs. In Graph A, K increases with frequency — correct. In Graph C, K increases with frequency — but curved — which is wrong. So between A and C, A is better. So **A** is the answer. I think I have to box it. So the answer is \boxed{A} But I am still unsure because it's not physically accurate. But given the options, it's the only reasonable choice. So I will submit \boxed{A} </think> <answer> A </answer> ✔

Figure F.6: A qualitative example on MMMU (Part II). Continuation of the MMMU case in Figure F.5. CGC better aligns the visual graph pattern with the required scientific relationship and finally selects the correct answer, whereas the base model follows a less reliable reasoning path.

MMStar



Question:
Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end. Question: Choose the answer.

Ground Truth: A

Qwen3-VL-8B:
<think> The question presents a visual analogy: "Is to As Is to", where the first pair of shapes (left) is compared to the second pair (right), and we must find the correct answer for the right side. Looking at the left side: - The first shape is a 3D cube-like object with a notch or indentation on one face. - The second shape is a 2D projection or unfolded net of that 3D object, showing how the faces are connected. Now, looking at the right side: - The first shape is a 3D cube-like object with a notch or indentation on one face — this is identical in form to the first shape on the left. - The second shape (the blank) should be the 2D projection or unfolded net corresponding to this 3D object. Examining the options: - A: Shows a 3D object with a notch — this is the same as the first shape on the right, not the net. - B: Shows a 3D object with a notch — again, this is the same as the first shape, not the net. - C: Shows a 2D net — this looks like the unfolded net of a cube with a notch, matching the pattern of the left side. - D: Shows a 3D object — not the net. - E: Shows a 3D object — not the net. Therefore, the correct answer is C, as it matches the pattern: 3D object → its 2D net. </think> <answer> C </answer> ✗

CGC-8B:
<think> The question presents a visual analogy: "Is to As Is to" — meaning the relationship between the first two shapes should be the same as the relationship between the third shape and the answer choice. Let's analyze: **Left pair:** - First shape (left): A 3D cube with a notch or indentation on one of its visible faces. - Second shape (middle): The same cube, but now viewed from a different angle — specifically, a perspective that shows more of the top and side, and the notch is now visible in a different orientation. It's essentially a **rotation** or change in viewpoint. **Right pair:** - Third shape (right): A 3D cube with a notch on one face, similar to the first shape. - We need to find the shape that results from rotating this cube to a new perspective — just like the second shape was a rotated view of the first. Now examine the options: - **A**: Shows the cube with a notch, but from an angle that makes the notch appear on a different face than in the original. This is plausible — it's a rotation. - **B**: Shows the cube with a notch, but from a front-facing angle — similar to the original. This doesn't show a significant change in perspective. - **C**: Shows the cube without any notch — this breaks the pattern, since the notch is a key feature. - **D**: Shows the cube with a notch, but from a different angle — perhaps a side view. This is also plausible. - **E**: Shows the cube with a notch, but from a very different perspective — almost a top-down view. This is also possible. But we must match the *exact* transformation. Looking at the left pair again: - The first shape has the notch on what appears to be the front-right face. - The second shape has the notch on the front-left face — so it's a **90-degree rotation around the vertical axis**. Now look at the right pair: - The third shape has the notch on the front-right face. - We need to rotate it 90 degrees around the vertical axis to get the notch on the front-left face — which would be option **A**. Wait — let me check the orientations more carefully. Actually, looking closely at the second shape in the left pair, it's not exactly a 90-degree rotation — it's a **perspective change** that reveals a different face. But the key is: the notch is still present, and its position relative to the cube's edges changes. In the right pair, the third shape has the notch on the front-right face. If we rotate it 90 degrees clockwise around the vertical axis, the notch will move to the front-left face — which matches option **A**. Option **D** also looks like a rotated version — but let's compare the notch's position. Looking at the original third shape (rightmost in the top row), the notch is on the bottom-right corner of the front face. In option **A**, the notch is on the bottom-left corner of the front face — which is a 90-degree rotation. In option **D**, the notch is on the top-right corner — which is a different rotation. So **A** is the correct match for the transformation shown in the left pair. Therefore, the correct answer is: **A** </think> <answer> A </answer> ✓

Figure F.7: A qualitative example on MMStar. The task requires solving a visual analogy under 3D structural transformation. CGC correctly captures the transformation pattern and selects the right option, whereas the base model is distracted by partial local similarity and predicts an incorrect answer.