MiCo: Multi-image Contrast for Reinforcement Visual Reasoning

Xi Chen[1] **Mingkang Zhu**[3] **Shaoteng Liu**[3] **Xiaoyang Wu**[1] **Xiaogang Xu**[3]
Yu Liu[2] **Xiang Bai**[4] **Hengshuang Zhao**[1*]

[1]HKU [2] Tongyi Lab, Alibaba Group [3]CUHK [4]HUST

Abstract

This work explores enabling Chain-of-Thought (CoT) reasoning to link visual cues across multiple images. A straightforward solution is to adapt rule-based reinforcement learning for Vision-Language Models (VLMs). However, such methods typically rely on manually curated question-answer pairs, which can be particularly challenging when dealing with fine-grained visual details and complex logic across images. Inspired by self-supervised visual representation learning, we observe that images contain inherent constraints that can serve as supervision. Based on this insight, we construct image triplets comprising two augmented views of the same image and a third, similar but distinct image. During training, the model is prompted to generate a reasoning process to compare these images (*i.e.*, determine same or different). Then we optimize the model with rule-based reinforcement learning. Due to the high visual similarity and the presence of augmentations, the model must attend to subtle visual changes and perform logical reasoning to succeed. Experiments show that, although trained solely on visual comparison tasks, the learned reasoning ability generalizes effectively to a wide range of questions. Without relying on any human-annotated question-answer pairs, our method achieves significant improvements on multi-image reasoning benchmarks and shows strong performance on general vision tasks.

1 Introduction

Making visual analysis with multiple images is crucial in many real-world applications. For example, we understand actions through sequential images or videos, gain 3D awareness by recognizing multiview images, and analyze events by observing differences between states, *etc*. Although Vision Language Models (VLMs) [2, 8, 19, 15, 1] demonstrate promising capabilities in understanding single images, we find them struggle to link visual cues across multiple images.

Multi-image understanding requires not only identifying fine-grained visual cues but also performing logical reasoning to uncover correspondences and differences among images. Recently, reasoning in language models [12, 16, 31, 27] has been significantly improved through the use of Chain-of-Thought (CoT) prompting, especially when combined with rule-based reinforcement learning [28]. Therefore, a straightforward idea to improve multi-image understanding is to extend this reinforcement learning paradigm to the visual domain. However, GRPO [28] requires constructing question-answer pairs with standard answers to compute rewards, which is particularly challenging for tasks involving fine-grained visual details and complex logic across images.

Instead of focusing on constructing QA pairs, we explore how to incentivize VLMs to perform multi-image reasoning with minimal data preparation cost. Modern VLMs already possess strong perceptual and multimodal capabilities. Meanwhile, recent advances in RL-based single image

*Corresponding Author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

reasoning [34, 23, 5] suggest that reasoning ability can be effectively acquired with limited data. However, most of these methods still rely on task-specific supervision, such as hand-crafted QA pairs. To reduce the reliance on manual annotations, we draw inspiration from self-supervised visual representation learning [7, 14, 13, 3], where images are used as their own source of supervision. Contrastive learning methods [14, 7], for instance, learn discriminative representations by pulling together features from different views of the same image and pushing them away from those of different images. Guided by this principle, we exploit inherent constraints in images as a supervision signal for reward calculation, and present a novel method, `MiCo` (**M**ultiple **i**mage **Con**trast).

Specifically, we construct training triplets consisting of two augmentations of the same image and a third, different but similar image with its own augmentation. We prompt the VLM to output the thinking process and make comparisons among these images to answer same/different. Multiple trajectories are sampled per example, and reinforcement learning is applied using advantages computed from the correctness of the final answer. A key aspect of our approach is the design of challenging image comparisons. If negative samples are too distinct, the reasoning is trivial. We address this by sampling frames from the same video or using image editing datasets to find similar images, ensuring subtle differences that require careful visual inspection and reasoning. Beyond this contrastive framework, we also introduce *Augmented GRPO*, a training strategy that samples trajectories using weak augmentations and optimizes them under stronger augmentations. This design allows high-quality CoTs to generalize to more difficult images.

Although the model is trained solely on the image comparison task, the learned ability to link visual cues across multiple images generalizes to a wider scope of scenarios. For example, the model can predict plausible future actions by analyzing visual changes across frames, distinguish object identities by comparing fine-grained appearance details, or detect subtle camera movement in scene transformations. Moreover, the contrastive learning process encourages attention to fine-grained details, which also benefits certain single-image understanding tasks like fine-grained layout/attribute understanding. Experimental results show that `MiCo` achieves strong performance for multi-image understanding [40, 32, 10], and also brings improvements on general vision tasks [11, 6, 39].

2 Related Work

Vision language model reasoning. Recent studies show that reasoning-capable LLMs [16, 12, 31, 27] can be effectively guided to generate long CoT [36] reasoning processes through reinforcement learning, leading to significant progress on tasks involving complex logic. Building on these advances, a surge of recent works [5, 29, 26, 25] has extended CoT reasoning into the vision-language domain. For example, MM-Eureka [23] expands training data coverage across domains and refines RL training strategies. NoisyRollout [20] introduces image augmentations to enrich the exploration space for policy optimization. LVAA-Thinking [4] provides a detailed analysis of supervised fine-tuning and RL for visual reasoning, along with a curated dataset for related tasks. ThinkLite [34] further improves data efficiency via sample selection with Monte Carlo Tree Search. While these methods rely heavily on curated training data generated by existing models or human annotations, our work explores an alternative: leveraging inherent constraints within visual data to naturally elicit reasoning ability—without explicit question-answer supervision.

Multi-image understanding. Understanding multiple images is crucial in real-world scenarios that require comparing object states, tracking actions, or recognizing objects across views. Recent large VLMs [2, 18, 17, 1, 15] have begun to support multi-image inputs natively. LLaVA-Interleave [19] extends LLaVA [19] to process interleaved multimodal inputs. VISC [41] introduces focus-centric data to enhance visual reasoning. Meanwhile, numerous benchmarks [32, 10, 24, 21, 44, 37] have been proposed to evaluate multi-image understanding from various angles. Despite these developments, recent evaluations [40, 46] highlight persistent limitations: VLMs often fail to link fine-grained visual cues across images, such as identifying the same object under different views or detecting subtle state changes for predictive reasoning. Our work addresses this gap by incentivizing the model to compare the fine details across images and make logical analysis.

3 Method

3.1 Pilot Study for Multi-image Understanding

We begin with a pilot study to assess how well current VLMs understand multiple images. As shown in Fig. 1, we present examples that highlight the capabilities of several state-of-the-art VLMs, Qwen2.5-

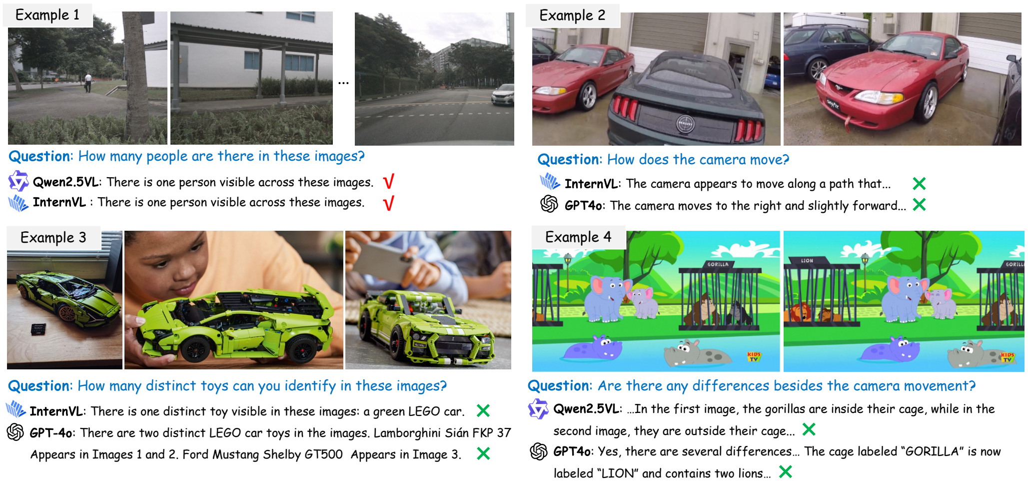

Figure 1: **Challenges for multi-image understanding.** While recent works support multiple images as input, most of them focus on scenarios where each image can be interpreted independently (*e.g.*, Example 1), which remains relatively easy for current state-of-the-art VLMs. However, many real-world tasks (*e.g.*, Example 2-4) require models to compare subtle visual differences, align visual cues across images, and reason about object correspondences—capabilities that current VLMs still struggle with. We gather Example 1 from MuirBench [32], Example 3 from VLM2-bench [40], Example 2,4 from real world samples.

VL [33], InternVL [8], and GPT-4o [15]. While many recent models and benchmarks [24, 32, 44] support multi-image or video inputs, they primarily focus on scenarios like Example 1, where each image can be understood in isolation. In Example 1, models correctly identify a single person across images, reflecting solid basic perception. However, when we examine more complex cases in Fig. 1, we observe that VLMs often suffer from severe hallucinations. In Example 2, both models fail to infer the correct camera movement, showing weaknesses in spatial reasoning. In Example 3, VLMs cannot distinguish between different car toys, indicating difficulty with cross-image comparison. Example 4 further reveals failure in tracking semantic changes across images, with hallucinated object positions and misidentified labels. These cases highlight that current VLMs still lack robust visual comparison abilities essential for multi-image understanding. These examples typically require the model to explicitly link visual cues across images, analyze fine-grained differences, and reason about inter-image correspondences.

As current VLMs already possess strong abilities in single-image perception (*e.g.*, reading fine-grained text) and demonstrate solid commonsense knowledge, as evidenced by their performance on standard vision benchmarks. We hypothesize that their primary limitation in multi-image understanding lies in their inability to compare and connect visual information across images. To address this gap, we focus on enhancing the meta-cognitive ability of **visual comparison**, the core skill needed for effective multi-image reasoning.

3.2 Multi-image Contrast

Rather than collecting data for each specific multi-image task, we aim to improve VLMs' general capacity to analyze and reason over multiple images by targeting the core meta skill: **visual comparison**. Inspired by the principles of self-supervised learning, we design a lightweight and scalable framework that encourages the model to distinguish similar yet distinct images. By simulating contrastive visual situations and prompting the model to generate structured reasoning trajectories, we aim to enhance its ability to perceive fine-grained differences, establish correspondences, and perform step-by-step comparisons across images.

Here, we elaborate on the pipeline of `MiCo`. The overall framework consists of the following main steps. First, we identify and construct contrastive image samples that are visually similar yet different. Then, we apply data augmentation to build informative training triplets. Finally, we leverage Augmented GRPO to evaluate a set of reasoning trajectories and optimize the VLM accordingly.

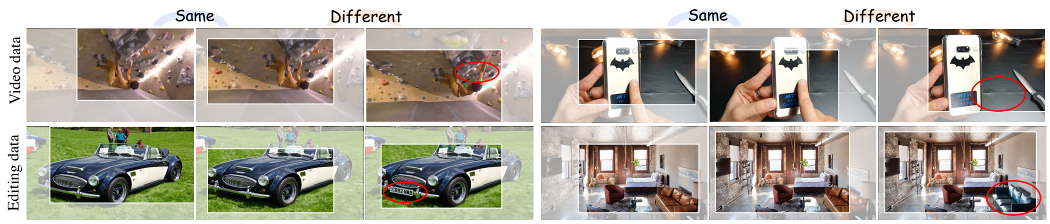

Figure 2: **Demonstrations for contrastive samples.** The first row shows two triplets from the video, and the second row demonstrates samples from image editing datasets. These samples are visually similar but contain subtle differences (marked with red circles), on which we apply random cropping and resizing. In each triplet, the first two images are the same, and the third image is different.

Image selection. We begin by selecting image pairs that are visually similar but exhibit subtle differences, which serve as contrastive supervision signals. We denote such a pair as (I_a, I_b), where I_a and I_b are distinct images sharing high structural similarity (*e.g.*, similar layout or background), but with small detail variations.

We leverage two types of data sources that naturally fulfill this requirement: video frames and image editing datasets. For video data, we randomly sample (I_a, I_b) from the same video with a temporal gap of 2 seconds, and compute their Structural Similarity (SSIM) to filter out near-identical pairs. For image editing data, each (I_a, I_b) pair consists of a "before" and "after" edited image. We compute the pixel-wise Mean Squared Error (MSE) to remove significantly different pairs. These constraints ensure that the collected pairs exhibit subtle but meaningful changes.

Image augmentation. While the selected image pairs already exhibit subtle variations, directly learning to distinguish them may still lead to shortcut learning. To increase task complexity and encourage detailed reasoning, we apply data augmentation to create diverse image views.

As the visualization examples provided in Fig. 2, given a source image I, we generate two augmented versions via random cropping and resizing (they do not change the content of images). For each image pair (I_a, I_b), we thus construct a contrastive triplet:

$$\mathcal{T} = \{\mathcal{T}_1(I_a), \mathcal{T}_2(I_a), \mathcal{T}_3(I_b)\},$$

3.3 Augmented GRPO

Question-answer formulation. After getting the triplets that contain similar images and their augmentations. We construct QA pairs for reinforcement learning. Given an image triplet, we add reasnoning prompt and user questions as follows:

Reasoning Template of MiCo

Reasoning Prompt: First output the thinking process in <think> </think> and give the final answer in <answer> </answer> tags.

User Question: Regardless of the augmentation, are image1 and image2 the same? How about image2 and image3, image1 and image3? Only return T(True) or F(False) in <answer> </answer>, for example <think> </think> <answer>TFT</answer>.

To increase the diversity and balance the difficulties of questions, besides the image triplet, we also construct image pairs and design the corresponding prompts for comparing two images. In addition, we use GPT-4o [15] to expand the user question of the same meaning but with various expressions.

Rollout augmentation. For each question q with augmented images $\mathcal{T}$, the original GRPO samples a group of outputs $\{o_1, o_2, \cdots, o_G\}$ from the old policy $\pi_{\theta_{old}}$ and then optimizes the policy model π_θ. To better leverage difficult samples that arise from strong augmentations, we sample trajectories using weakly augmented inputs $\mathcal{T}^w$, which are easier to produce valid reasoning chains. These sampled trajectories are then used to optimize the policy on stronger augmented contexts $\mathcal{T}^s$, effectively transferring reliable behavior to harder instances.

Algorithm 1 `MiCo`: Reinforcement Multi-image Reasoning

Input: Policy π_θ, old policy $\pi_{\theta_{\text{old}}}$, image triplet dataset $\mathcal{D} = \{(I_1, I_2, I_3)\}$, training steps T_{max}, group size G, clip parameter ϵ, weak augment operators $\mathcal{T}^{\text{w}}$, strong augment operators $\mathcal{T}^{\text{s}}$
for $t = 1$ to T_{max} **do**
  Sample triplet $(I_1, I_2, I_3) \sim \mathcal{D}$
  Apply weak augmentation: $(I_1^w, I_2^w, I_3^w) = \mathcal{T}^{\text{w}}(I_1, I_2, I_3)$
  Apply strong augmentation: $(I_1^s, I_2^s, I_3^s) = \mathcal{T}^{\text{s}}(I_1, I_2, I_3)$
  Construct prompts $\mathbf{q}^w$ and $\mathbf{q}^s$ from the weak and strong augmented triplets, respectively
  Sample G CoT responses $\{\mathbf{o}_i\}_{i=1}^G$ from $\pi_{\theta_{\text{old}}}(\cdot \mid \mathbf{q}^w)$ ▷ Rollouts from weak prompt
  Evaluate reward $R_i = R(I^w, \mathbf{q}^w, \mathbf{o}_i)$ for each $i = 1, \ldots, G$
  Compute group baseline $\bar{R} = \frac{1}{G}\sum_{i=1}^G R_i$, and advantages $\hat{A}_i = \frac{R_i - \bar{R}}{\sigma(R)}$
  Optimize π_θ on the strong prompt $\mathbf{q}^s$ using the group rollouts:
  $L(\theta) = \frac{1}{G}\sum_{i=1}^G \min\left(r_i \hat{A}_i, \text{clip}(r_i, 1-\epsilon, 1+\epsilon)\hat{A}_i\right)$, where $r_i = \frac{\pi_\theta(\mathbf{o}_i|\mathbf{q}^s)}{\pi_{\theta_{\text{old}}}(\mathbf{o}_i|\mathbf{q}^s)}$
  $\theta \leftarrow \theta - \nabla_\theta L(\theta)$
  $\theta_{\text{old}} \leftarrow \theta$
end for

Training objective. The training objective of Augmented GRPO could be formulated as follows. This objective encourages the policy to assign higher likelihoods to responses with higher relative rewards within each group.

$$\begin{aligned}\mathcal{J}(\theta) &= \mathbb{E}[q \sim P(Q,\ \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)] \\ &= \frac{1}{G}\sum_{i=1}^G \left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i,\ \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}\right)A_i\right) - \beta\,\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right),\end{aligned} \tag{1}$$

$$\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right) = \frac{\pi_{ref}(o_i \mid q)}{\pi_\theta(o_i \mid q)} - \log\frac{\pi_{ref}(o_i \mid q)}{\pi_\theta(o_i \mid q)} - 1, \tag{2}$$

where ϵ and β are hyperparameters. Following GRPO [28], A_i is the normalized advantage computed based on rewards $\{r_1, r_2, \ldots, r_G\}$.

$$A_i = \frac{r_i - mean(\{r_1, r_2, \cdots, r_G\})}{std(\{r_1, r_2, \cdots, r_G\})}. \tag{3}$$

Following DeepSeek-R1 [12], we leverage the binary format reward and the accuracy reward, which considers the matching of "<think> </think> <answer> </answer>" tags, and the correctness of the final answer. For the triplet comparisons, we get 1 for the accuracy reward only if we make correct comparisons for all three pairs.

Overall algorithm. Our `MiCo` could be summarized in Algorithm 1. We first construct an image triplet consisting of two augmented views of the same image and a third, visually similar but distinct image (with augmentations). The model is prompted to perform multi-image comparison and generate reasoning trajectories. During training, chain-of-thought responses are sampled from the weakly augmented views, and the policy is optimized on the strongly augmented ones using rule-based reinforcement learning. This process enables the model to learn fine-grained visual reasoning in a self-supervised manner.

4 Experiments

4.1 Implementation Details

Hyper-parameters. For the baseline model, we follow previous works [23, 20, 34, 4] and select Qwen2.5-VL-7B [2]. For the training data, we use OmniEdit [35] for image editing pairs and extract video frames from Vidgen-1M [30]. The part of reinforcement learning follows GRPO [28], we set a

Table 1: **Performance on VLM2-Bench** [40], which evaluates the ability to compare and link fine-grained visual cues across multiple images. Without relying on any human- or model-annotated data, `MiCo` achieves significant improvements and sets a new state-of-the-art. Reasoning-based models (marked with ●) are evaluated using their corresponding prompting strategies.

Baselines or Models	General		Object			Person				Overall*	
	Mat	Trk	Cpr	Cnt	Grp	Cpr	Cnt	Grp	VID	Avg	Δ_{human}
Chance-Level	25.00	25.00	50.00	34.88	25.00	50.00	34.87	25.00	-	32.73	-61.44
Human-Level	95.06	98.11	96.02	94.23	91.29	97.08	92.87	91.17	100.00	95.16	0.00
○ LLaVA-OneVision[17]	16.60	13.70	47.22	56.17	27.50	62.00	46.67	37.00	47.25	39.35	-55.81
○ LLaVA-Video-7B [43]	18.53	12.79	54.72	62.47	28.50	62.00	66.91	25.00	59.00	45.65	-49.51
○ LongVQA-7B [42]	14.29	12.98	46.53	49.47	29.00	58.00	41.56	25.00	45.00	37.10	-58.06
○ mPLUG-Owl2-7B [38]	17.37	18.26	49.17	62.97	31.00	63.00	58.06	29.00	43.00	40.87	-54.31
○ Qwen2-VL-7B [2]	18.07	19.18	68.08	61.84	37.50	72.00	67.92	47.00	55.25	49.76	-45.40
○ InternVL2.5-8B [8]	41.24	26.53	72.22	67.65	40.00	**85.00**	66.67	52.25	50.25	55.41	-39.75
○ InternVL2.5-26B [8]	30.50	30.59	43.33	51.48	52.50	59.50	59.67	61.25	45.25	45.59	-49.57
○ Qwen2.5-VL-7B [2]	35.91	43.38	71.39	41.72	47.50	80.00	59.76	69.00	45.00	54.82	-40.34
○ GPT-4o [15]	37.45	39.27	74.17	**80.62**	57.50	50.00	**90.50**	47.00	**66.75**	60.36	-34.80
● MM-Eureka-7B [23]	55.60	47.03	74.10	52.50	54.00	77.50	60.00	51.00	43.50	57.24	-37.91
● NoisyRollout-7B [20]	40.93	43.83	63.33	50.83	34.50	70.50	63.33	47.00	36.50	50.08	-45.08
● ThinkLite-VL-7B [34]	40.45	46.58	75.56	62.50	49.50	77.50	62.50	51.00	36.50	55.79	-39.37
● VLAA-Thinker-7B [4]	47.49	63.03	72.20	61.40	55.00	71.00	57.50	51.00	47.75	58.49	-36.67
○ Qwen2.5-VL-7B-CoT[2]	43.24	42.92	66.39	50.56	36.00	62.50	55.83	39.00	36.75	48.91	-46.24
● `MiCo`-7B-CoT	**57.14**	**67.12**	**81.94**	56.67	**58.00**	65.00	57.50	62.00	44.25	**61.06**	**-34.09**
Δ Improvement	+13.90	+24.20	+15.55	+6.11	+22.00	+2.50	+1.67	+23.00	+7.50	+12.93	+12.93

format reward and accuracy reward with the weight of 1:1, respectively. Besides, we also apply a KL regularization with a weight of 0.01. During training, we follow previous works [23] to skip the rollout group with all correct/false answers. During training, we use a learning rate of 1e-6 and set the batch size of 16. For each training sample, we generate a group of 8 rollouts. We train the model for 600 iterations on 8×A100 GPUs.

Evaluation protocols. For evaluation, we follow the default hyper-parameters of Qwen2.5-VL [33] and utilize the VLMEvalKit [9]. For reasoning baselines, we adopt their official prompting formats. Minor inconsistencies in results may occur due to differences in the implementation details of evaluation frameworks or answer parsing logic.

4.2 Result Analysis for Multi-image Comparison

Evaluation metrics. We first report the model performance on VLM2-Bench [40]. This benchmark mostly aligns with our intention of linking fine-grained visual cues across images. Specifically, VLM2-Bench includes three tracks: General Cue (GC), Object-centric Cue (OC), and Person-centric Cue (PC). Each track consists of subtasks with specific metrics: Mat (Matching) and Trk (Tracking) use paired T/F accuracy; Cpr (Comparison) evaluates consistency by requiring the model to correctly answer both a positive and its corresponding negative statement; Cnt (Counting) uses normalized error to measure numerical prediction accuracy; Grp (Grouping) is a multiple-choice task assessing clustering ability; and VID (Video Identity Describing) is scored based on GPT-4o evaluation of open-ended descriptions.

Result analysis. As shown in Tab. 1, we present the comparison results on VLM2-Bench. We observe that all existing open- and closed-source models lag behind human performance by a large margin. Among them, GPT-4o [15] demonstrated clear advantages over other models. Thanks to the strong generalization ability of reinforcement learning, recent reasoning VLMs [23, 20, 34, 4] have shown consistent improvements when built upon Qwen2.5-VL-7B [2]. We report the performance of `MiCo` in the final block. Trained with contrastive triplets, `MiCo` effectively learns the core ability to compare images, achieving substantial gains across multiple tasks and obtaining the best average performance overall. Notably, our 7B model even outperforms GPT-4o. However, we find that CoT reasoning does not benefit all sub-tasks equally. Specifically, for tasks involving human faces (person track), CoT-based models offer limited or even negative gains compared to no-CoT counterparts. We hypothesize that human identity representations, such as facial nuances, are difficult to verbalize, thus limiting the benefit of language-based reasoning. In contrast, object-level identity differences (*e.g.*, logos,

Table 2: **Ablation studies** on key configurations. We conduct experiments on VLM2-Bench [40] and report the average accuracy across the general, object, and person tracks. For each ablation, all other settings are kept consistent with our final model to ensure fair comparisons.

(a) Learning Paradigm

	General	Object	Person
Qwen2.5-VL [2]	39.64	53.53	63.44
SFT	42.90	51.15	55.98
No-CoT RL	45.36	50.01	55.23
CoT RL	62.13	65.53	57.18

(b) Data Source

	General	Object	Person
Edit Data[1] [35]	61.23	65.33	56.35
Edit Data[2] [45]	60.88	64.33	55.27
Video Data	60.29	64.50	55.68
Edit[1] + Video	62.13	65.53	57.18

(c) Rollout Augmentation

	General	Object	Person
Qwen2.5-VL [2]	39.64	53.53	63.44
(Strong, Strong)	59.41	64.00	56.98
(Weak, Weak)	55.58	62.03	54.81
(Weak, Strong)	62.13	65.53	57.18

(d) Sample Formulation

	General	Object	Person
Qwen2.5-VL [2]	39.64	53.53	63.44
Image Pairs	56.41	66.98	55.68
Image Triplets	60.64	65.33	55.81
Pairs + Triplets	62.13	65.53	57.18

(e) Prompt Diversity

	General	Object	Person
Qwen2.5-VL [2]	39.64	53.53	63.44
Single Prompt	55.53	64.16	51.50
20 Variations	62.13	65.53	57.18
50 Variations	63.13	65.29	54.93

(f) Image Augmentations

	General	Object	Person
Base (Crop, Resize)	62.13	65.53	57.18
Base + Flip	61.13	63.98	56.77
Base + Rotat.	62.58	65.03	55.86
Base + Color.	60.15	64.26	54.86

textures, shapes) are more readily describable, allowing CoT reasoning to help reduce hallucinations and improve distinction.

4.3 Ablation Studies

We conduct a series of ablation studies to validate the effectiveness of our core designs. As an initial exploration of visual reasoning, we also analyze the impact of some basic configurations.

Training strategies. In Tab. 2 (a), we evaluate different training paradigms. We first apply supervised fine-tuning (SFT) on our contrastive dataset, allowing the model to directly predict the final answer. We observe that this leads to minor gains on the general track, which more closely aligns with the training task. However, the ability acquired through SFT does not generalize well to more diverse reasoning tasks. We also test a no-CoT reinforcement learning baseline, where the model is trained to output answers directly using GRPO [28]. Due to the absence of intermediate reasoning steps, the resulting trajectories are short and behave similarly to SFT, yielding limited improvements.

Data source. In Tab. 2 (b), we compare different training data sources. Both the image editing data (OmniEdit [35]) and video-derived frames (VidGen [30]) individually support effective learning. Combining these two heterogeneous sources further enhances performance. We also validate that our framework is not tied to specific editing styles, as models trained on either OmniEdit or UltraEdit [45] generalize well, demonstrating robustness to the editing domain variation.

Rollout augmentation. In this work, we leverage weak augmentations for rollout sampling, and use these high-quality answers to optimize harder questions with stronger augmentations. In Tab. 2 (c), we report different combinations of augmentation in "(sampling, optimization)" process. We show that, strong augmentations are vital for contrastive learning compared with weak augmentations, and our rollout augmentation strategy gets the best performance.

Sample formulation. As discussed in Sec. 3.3, we construct prompts based on either image pairs or image triplets. While we initially suspected that binary image-pair comparisons (with 50% guess probability) might result in low-quality CoTs, our experiments reveal that they still contribute positively to performance. In practice, we find that combining both formats—pair-based and triplet-based leads to the best results.

Other configurations. In Tab. 2 (e), (f), we explore the effects of prompt and augmentation diversity. We observe that increasing the variation of image prompts helps prevent overfitting, with performance

Table 3: **Quantitative results on general vision benchmarks.** We report performance for wide scenarios. Multi-image benchmarks are marked in bold. `MiCo` brings steady improvements compared with our baseline, and gets competitive results against other visual reasoning models.

	MuirBench [32]	**BLINK** [10]	Hallusion [11]	MMStar [6]	MMMU [39]	MathVistas [22]
MM-Eureka-7B [23]	60.57	54.39	68.45	65.73	54.11	72.00
NoisyRollout-7B [20]	59.61	56.07	66.66	65.66	54.55	71.60
VLAA-Thinker-7B [4]	61.00	54.81	69.08	63.60	54.44	70.80
ThinkLite-VL-7B [34]	57.62	55.81	72.97	66.80	53.55	71.89
Qwen2.5VL-7B [2]	58.43	55.54	69.50	64.06	54.11	67.10
`MiCo`-7B	60.53	57.23	69.61	65.60	54.77	67.90
Δ Improvement	+2.10	+1.69	+0.11	+1.54	+0.66	+0.80

Table 4: **Task analysis for visual reasoning.** We list representative sub-tasks from MuirBench [32] and BLINK [10] to analyze the generalization ability and limitations for `MiCo`.

	Visual retrieval	Semantic Corr.	Spatial Rela.	Scene Under.	Forensic Det.	Relative Depth
Qwen2.5VL-7B [4]	63.69	33.09	88.81	61.82	48.48	81.45
MM-Eureka-7B [23]	57.19 +	33.09 +	82.51 -	67.74 ++	50.00 +	75.80 -
VLAA-Thinker-7B [4]	68.83 +	34.53 +	86.71 -	69.89 ++	47.72 -	76.61 -
`MiCo`-7B	71.23 ++	41.72 ++	90.20 +	63.97 +	47.72 -	78.22 -

saturating at around 50 distinct prompt templates. For image augmentations, we experimented with various techniques and ultimately selected random cropping and resizing as the default setting based on empirical performance.

4.4 Analysis on General Vision Tasks

In this section, we evaluate the generalization ability and capacity boundaries of `MiCo` on a broader range of vision tasks. We first report quantitative results on additional benchmarks and analyze performance across more diverse task types.

Results on additional benchmarks. As shown in Tab. 4, we evaluate `MiCo` on MuirBench [32] and BLINK [10], both of which are representative multi-image understanding benchmarks. To further assess generalization, we also include several single-image benchmarks, including MMStar [6], MMMU [39], HallusionBench [11], and MathVista [22].

Compared to methods trained with manually curated supervision, our contrastive learning framework exhibits strong performance on multi-image understanding tasks, where relational reasoning across images is crucial. While `MiCo` also improves over standard baselines on single-image tasks, its performance remains behind models trained with task-specific guidance, particularly in complex scenarios like visual mathematics, where symbolic reasoning and structured representation are essential but not explicitly modeled in our current training paradigm.

Task-wise analysis. We further analyze the performance of `MiCo` on specific sub-tasks from MuirBench [32] and BLINK [10] to better understand its strengths and limitations. Our contrastive learning framework demonstrates clear advantages on correspondence-style tasks, such as *Visual Retrieval* and *Semantic Correspondence*, where `MiCo` outperforms other reasoning models. These results highlight the model's strength in aligning multimodal signals through relational comparisons. In addition, *Spatial Relation* tasks—which evaluate the model's understanding of image layout and object positioning—also benefit from contrastive training. By encouraging attention to relative positions among visual entities, `MiCo` achieves the highest accuracy in this category.

On the other hand, `MiCo` lags behind models trained with manually curated reasoning datasets on tasks such as *Scene Understanding* and *Forensic Detection*, which typically rely on single-image question answering. These tasks often demand domain-specific priors or curated logic patterns that are less emphasized in our data construction process. We also observe that *Relative Depth*, a representative spatial reasoning task, remains challenging for all models and exhibits a noticeable performance drop. This suggests that depth-aware understanding is not sufficiently captured by current training signals. Incorporating explicit spatial or geometric cues remains a promising direction for future work toward building spatially grounded vision-language models.

Figure 3: **Demonstrations for visual reasoning.** Given a question, `MiCo` first examines the details of each image to identify answer-related visual cues, and then performs cross-image comparisons to derive the final answer. The reasoning processes are marked in gray, with key contents underlined.

4.5 Qualitative Analysis

We provide qualitative demonstrations in Fig. 3 to illustrate the multi-image reasoning process of `MiCo`. For each question, the model first generates a detailed reasoning chain that carefully compares visual cues across the input images. This includes analyzing fine-grained differences and similarities that are relevant to the question. Based on this reasoning, the model then produces a final answer that successfuly addresses the query objective.

4.6 Unsuccessful Attempts

Throughout our exploration, we experimented with several alternative approaches that ultimately did not lead to improved performance. For completeness and to facilitate future research, we summarize these unsuccessful attempts and provide insights into why they may have failed.

Confidence reweighting. Since our task is formulated as answering T/F questions, even when evaluating three comparisons simultaneously, there remains a non-trivial chance (12.5%) of obtaining the correct answer purely by guessing. To reduce the impact of such randomness, we explored adding an additional reward or weight based on the model's answer confidence. Specifically, we experimented with several approaches to compute confidence scores from the softmax probabilities of the output tokens. However, these confidence-based reweighting strategies did not yield any performance improvements. We analyze that this may be due to the fact that the softmax probability of the predicted token does not reliably reflect the model's true certainty about the overall answer. In particular, the model may assign high confidence to tokens that are syntactically or semantically unrelated to the actual correctness of the reasoning (e.g., punctuation, or irrelevant words within the output). As a result, the computed "confidence" can be misleading, making it an ineffective signal for reward shaping.

Importance sampling. As in our Augmented GRPO, we sample the trajectories on simple examples with weak augmentations, but we use the trajectory to optimize harder exaples with strong augmentations. This might cause misalignment similar to offline reinforcement learning. In this way, we apply importance sampling, which calculates the probability gap between the trajectories for the simple and hard examples as a weight to reweight the reward/advantages. This strategy could not bring improvements. We suspect that although importance sampling is theoretically justified, it may interfere with the core optimization dynamics of GRPO. Specifically, GRPO relies on the relative ranking of trajectories within a group to compute structured advantages. Introducing importance weights—derived from distribution shifts—may distort this internal ranking or inject instability into the reward signals. Additionally, the token-level probability changes caused by visual augmentations can be noisy or poorly calibrated, making the computed importance weights unreliable in practice.

5 Conclusion

In this work, we propose a self-supervised framework that leverages inherent image constraints to incentivize multi-image reasoning in VLMs. We identify that the core challenge lies in linking visual cues across images. To address this, we adopt contrastive learning principles and construct image triplets for reinforcement training. To further enhance reasoning, we introduce Augmented GRPO, which samples rollouts from simpler examples and optimizes the model on harder ones. Although trained solely on image comparison tasks, our model generalizes well and achieves strong results across multiple benchmarks.

Broader Impact. MiCo explores a self-supervised and reinforcement learning-based approach to improve multi-image reasoning in vision-language models without relying on human-annotated question-answer pairs. By leveraging intrinsic visual constraints, such as consistency across augmented views and differences between similar images, MiCo significantly reduces the need for labor-intensive data curation. This has the potential to democratize the development of reasoning-capable AI systems, making them more accessible in low-resource settings or for underrepresented languages and domains where curated datasets are scarce.

However, as with any powerful vision-language technology, there is a risk of misuse, particularly in applications involving surveillance, misinformation, or unauthorized inference of user intent from visual data. MiCo's improved ability to perform fine-grained comparisons across images could be exploited in privacy-invading scenarios if deployed irresponsibly. To mitigate such risks, we advocate for deploying MiCo in alignment with responsible AI guidelines, ensuring transparency, consent, and clear boundaries in its application domains. In practice, this includes integrating robust sensitive content filtering, restricting deployment in high-stakes or privacy-sensitive scenarios, and establishing human-in-the-loop mechanisms for critical decision-making processes.

Limitations and future directions. While our approach supports general reasoning through visual comparisons, it shows limited effectiveness on specialized tasks such as face verification, visual math, and spatial understanding, where structured priors or domain-specific knowledge are required. In future work, we plan to explore more efficient data construction strategies tailored to these domains.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (No. 624B2124, 62441615, 62422606).

References

[1] 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:2303.08774*, 2023. 1, 2

[2] 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:2502.13923*, 2025. 1, 2, 5, 6, 7, 8

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2

[4] 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*, 2025. 2, 5, 6, 8

[5] Liang Chen, Lei Li, Haozhe Zhao, and Yifan Song. Vinci. r1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V, 2025. 2

[6] 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*, 2024. 2, 8

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 1, 3, 6

[9] 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 *ACMMM*, 2024. 6

[10] 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 *ECCV*, 2024. 2, 8

[11] 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. 2, 8

[12] 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:2501.12948*, 2025. 1, 2, 5

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv:2410.21276*, 2024. 1, 2, 3, 4, 6

[16] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv:2412.16720*, 2024. 1, 2

[17] 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. *TMLR*, 2025. 2, 6

[18] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv:2407.07895*, 2024. 2

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2

[20] 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:2504.13055*, 2025. 2, 5, 6, 8

[21] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. In *NeurIPS*, 2025. 2

[22] 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. In *ICLR*, 2024. 8

[23] 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:2503.07365*, 2025. 2, 5, 6, 8

[24] 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. In *ICLR*, 2025. 2, 3

[25] Yi Peng, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, et al. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. *arXiv:2504.05599*, 2025. 2

[26] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv:2503.07536*, 2025. 2

[27] ByteDance Seed, Yufeng Yuan, Yu Yue, Mingxuan Wang, Xiaochen Zuo, Jiaze Chen, Lin Yan, Wenyuan Xu, Chi Zhang, Xin Liu, et al. Seed-thinking-v1. 5: Advancing superb reasoning models with reinforcement learning. *arXiv:2504.13914*, 2025. 1, 2

[28] 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:2402.03300*, 2024. 1, 5, 7

[29] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv:2503.20752*, 2025. 2

[30] Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. Vidgen-1m: A large-scale dataset for text-to-video generation. *arXiv:2408.02629*, 2024. 5, 7

[31] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv:2501.12599*, 2025. 1, 2

[32] 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. In *ICLR*, 2025. 2, 3, 8

[33] 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:2409.12191*, 2024. 3, 6

[34] 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. 2, 5, 6, 8

[35] Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. In *ICLR*, 2024. 5, 7

[36] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. 2

[37] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. Towards open-ended visual quality comparison. In *ECCV*, 2024. 2

[38] 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*, 2024. 6

[39] 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. 2, 8

[40] 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*, 2025. 2, 3, 6, 7

[41] Juntian Zhang, Yuhan Liu, Wei Liu, Jian Luan, Rui Yan, et al. Weaving context across images: Improving vision-language models through focus-centric visual chains. *arXiv:2504.20199*, 2025. 2

[42] 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*, 2024. 6

[43] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv:2410.02713*, 2024. 6

[44] Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv:2406.12742*, 2024. 2, 3

[45] Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. In *NeurIPS*, 2024. 7

[46] 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. 2

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Please refer to introduction in Sec. 1.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Please refer to the limitations in Sec. 5.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [NA]

Justification: This paper does not include theoretical result.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Please refer to Sec. 3 and Sec. 4.

   Guidelines:

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

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

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

Answer: [No]

Justification: This work is cooperating with the company, we need to apply for approval.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Please refer to Sec. 4.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [NA]

   Justification: The results are stable, papers in related topic do not require error bars.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Please refer to Sec. 4.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Yes, it conforms to the requirements.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Please refer to the Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [NA]

Justification: The scope of our research does not involve data or models with a high risk for misuse. We utilize publicly available, non-sensitive datasets and models that do not require special safeguards for responsible release.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification:We have carefully credited all previous works we used in the paper. The license and terms are properly respected.

Guidelines:

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

13. **New assets**

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

Answer: [NA]

Justification: This question does not apply to our research as no new assets were introduced. Our study solely relies on existing, publicly available resources, and thus, no additional documentation for new assets is necessary.

Guidelines:

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

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

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

Answer: [NA]

Justification: This question does not apply to our research as it does not involve crowdsourcing experiments or research with human subjects.

Guidelines:

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

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

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

Answer: [NA]

Justification: This question is not applicable to our research as no human subjects were involved.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: We only use the LLM for writing, editing, or formatting purposes.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.