CR^3: Boosting Compositional Reasoning in MLLMs through Rule-based Reinforcement Learning

Shun Qian[1], **Bingquan Liu**[1], **Chengjie Sun**[1, 3], **Peijin Xie**[1], **Zhen Xu**[2], **Baoxun Wang**[2]
[1]Faculty of Computing, Harbin Institute of Technology
[2]Platform and Content Group, Tencent
[3]National Research Center for Language Technology and Digital Economy, Harbin Institute of Technology
shunqian@insun.hit.edu.cn, liubq@hit.edu.cn, cjsun@insun.hit.edu.cn, xpj@stu.hit.edu.cn,
zxu@insun.hit.edu.cn, asulewang@tencent.com

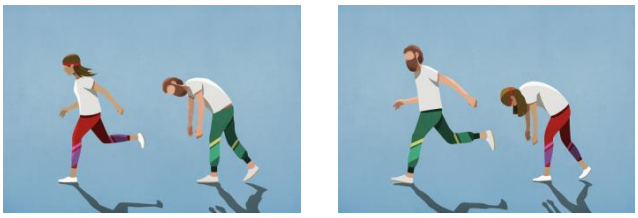
Abstract

Compositional reasoning is a critical capability for multimodal models, enabling systematic understanding of complex scenes through structured combinations of objects, attributes, and relations. However, existing research on this ability primarily focuses on vision-language models (VLMs, e.g., CLIP and SigLIP), with limited exploration of multimodal large language models (MLLMs). To address this gap, we introduce CR^3, a novel framework that enhances compositional reasoning abilities of MLLMs via rule-based reinforcement learning. CR^3 leverages rule-based rewards to optimize the MLLM's policy on systematically curated multimodal instruction-following tasks, guided by a model-adaptive dynamic task mixing strategy. Our approach boosts performance by over 19% on three compositional reasoning benchmarks, significantly outperforming supervised fine-tuning (SFT) by at least 12%. Crucially, CR^3 demonstrates superior generalization by improving performance on out-of-domain benchmarks where SFT methods degrade, highlighting its effectiveness and data efficiency.

Code — https://github.com/AAbathur/CR3
Datasets —
https://github.com/AAbathur/CR3/tree/main/train/data

1 Introduction

Compositional reasoning, the ability to decompose complex scenes or descriptions into interpretable elements (objects, attributes, relationships) and reconstruct through structured combinations, is a hallmark of human intelligence (Ma et al. 2023; Janssen and Partee 1997). However, most SoTA vision-language models (VLMs), such as CLIP and FLAVA, often behave as "bag-of-words" models due to their limited compositional reasoning. While these models excel at recognizing isolated objects, they frequently fail to bind attributes (e.g., color, size, shape) or states (e.g., eating, broken, standing) to objects, or to interpret spatial and logical relationships (e.g., above, below, left, right). For instance, given an image of "a red cube left of a metallic sphere", these models may assign similarly high confidence to the incorrect description "a metallic cube left of a red sphere", revealing a fundamental limitation in compositional reasoning.



Which image does the caption "the person with green legs is running quite slowly and the red legged one runs faster" best describe?

The caption "the person with green legs is running quite slowly and the red legged one runs faster" best describes the second image. In this image, the person with green legs appears to be running ahead and faster, while the one with red legs is trailing behind and looks fatigued.

Figure 1: An example from Winoground (Thrush et al. 2022) and GPT-4o's response on it. Text in red signifies the wrong choice. The text in green denotes the correct description on selected image.

This critical limitation persists even in the most advanced multimodal large language models (MLLMs). Recent studies (Ni et al. 2025; Chen et al. 2024a; Tong et al. 2024) reveal that even cutting-edge MLLMs, such as GPT-4V (Achiam et al. 2023), exhibit significant limitations in handling relational inversions and long-tail compositions. As shown in Figure 1, even a leading model like GPT-4o can fail in simple compositional reasoning queries, generating responses with clear logical contradictions. Despite its importance, systematically improving the compositional reasoning capabilities of MLLMs remains an open and critical challenge.

A straightforward approach to address this challenge in MLLMs is Supervised Fine-Tuning (SFT) on specifically designed multimodal instruction datasets. However, this method faces critical scalability challenges. Curating comprehensive, high-quality data for diverse compositional reasoning scenarios is labor intensive and expensive. Furthermore, models trained via SFT often overfit to the specific

patterns in the training data, exhibiting poor generalization to novel compositions.

To overcome these limitations, we propose CR^3, a novel framework that enhances **C**ompositional **R**easoning in MLLMs through **R**ule-based **R**einforcement learning. Our framework first selects high-quality compositionally-aware image-text pairs from an open-source dataset using a multimodal collaborative filtering mechanism. These pairs are then systematically transformed into three distinct and verifiable instruction-following tasks designed to strengthen compositional reasoning abilities. Subsequently, we employ rule-based reward functions inspired by Deepseek-R1 (Guo et al. 2025), to evaluate the reward score of the MLLM's responses. The MLLM's policy is subsequently optimized with these rewards through the Group Relative Policy Optimization (GRPO) algorithm. Finally, to maximize synergy between the tasks, we introduce a model-adaptive dynamic mixing strategy. This approach intelligently adjusts the proportion of different tasks during training based on the model's performance at various stages. Finally, we propose a model-adaptive dynamic mixing strategy to maximize synergy between the three tasks. It dynamically adjusts task proportions during training based on the model's performance.

Extensive experimental results demonstrate the effectiveness of our CR^3 framework. When applied to SoTA MLLMs such as Qwen2.5-VL (Bai et al. 2025) and InternVL3 (Zhu et al. 2025), CR^3 consistently achieves performance gains exceeding 19% across three challenging compositional reasoning benchmarks, regardless of model scale or architecture. Notably, CR^3 maintains a significant advantage over standard SFT approaches, delivering at least 12% improvement on every baseline model. We further evaluate out-of-domain generalization using popular multimodal benchmarks (e.g., MMMU (Yue et al. 2024) and MMB (Liu et al. 2024)). The results reveal that CR^3 substantially enhances baseline models' performance on generic vision-language tasks, while SFT methods suffer performance degradation. These results highlight CR^3's superior generalization and data efficiency. Our main contributions are summarized as follows:

- We propose the first framework to enhance compositional reasoning in MLLMs through rule-guided reinforcement learning, establishing a novel paradigm for this critical capability.
- We construct and publicly release a high-quality, compositionally-aware visual instruction-following dataset, which is specifically designed to advance MLLM research.
- Through extensive experiments, we demonstrate that the CR^3-enhanced model exhibits robust generalization in compositional reasoning across diverse multimodal benchmarks.

2 Related Work

2.1 Multimodal Compositionality

Although VLMs have achieved remarkable success across diverse multimodal tasks, they often lack robust compositional understanding and reasoning capabilities. Studies like NegCLIP (Yuksekgonul et al. 2023) show that SoTA VLMs (e.g., CLIP (Radford et al. 2021), FLAVA (Singh et al. 2022), X-VLM (Zeng, Zhang, and Li 2022)) behave like "bag-of-words" models, failing to capture relational, attributive, and positional dependencies. DAC (Doveh et al. 2023) identifies the low quality of web-crawled captions, a core training data source, as a critical bottleneck, Subsequent work (Stone et al. 2025) proposes automated caption refinement to enhance pretraining dataset density. Building on this, TripletCLIP (Patel et al. 2024), GMN (Sahin et al. 2024), and SPEC (Peng et al. 2024) synthesize hard negative images via perturbed captions, explicitly training VLMs to distinguish subtle compositional differences. However, these efforts focus on VLMs, with limited exploration in MLLMs.

2.2 Rule-based Reinforcement Learning for MLLMs

Reinforcement learning (RL) has enhanced the reasoning capabilities of LLMs, as demonstrated by OpenAI O1 (Jaech et al. 2024), Kimi 1.5 (Team et al. 2025), and DeepSeek-R1 (Guo et al. 2025). Inspired by these advances, the multimodal field has adapted rule-based RL techniques to MLLMs. A series of studies (Xie et al. 2025; Feng et al. 2025; Liu et al. 2025) have extended the rule-based reinforcement learning strategy to various multimodal tasks. Moreover, MM-Eureka (Meng et al. 2025) and VisualTinker-R1-Zero (Zhou et al. 2025) explore to reproduce the aha moment in multimodal reasoning tasks. R1-Onevision (Yang et al. 2025) and OpenVLThinker (Deng et al. 2025) utilize pure-text R1 models to address the lack of high-quality multimodal reasoning data. R1V (Peng et al. 2025) and R1-VL (Zhang et al. 2025) refine reasoning via iterative strategies. To our knowledge, CR^3 is the first rule-based RL method specifically designed for multimodal compositional reasoning, with large-scale evaluations confirming its efficacy.

3 Approach

3.1 GRPO with Rule-Based Rewards

To present the CR^3 approach, this section provides a concise overview of the GRPO algorithm with rule-based rewards for reinforcement learning training.

Compared to Proximal Policy Optimizatio (PPO) (Schulman et al. 2017), GRPO demonstrates greater computational efficiency by eliminating the need for an additional critic model. Instead, GRPO directly estimates the policy model by evaluating the relative quality of multiple candidate responses. For a given question q, GRPO first samples G responses $\{o_1, o_2, ..., o_G\}$ from the old policy model π_{old} and computes the corresponding rewards$\{r_1, r_2, ..., r_G\}$ with the reward model. The advantage of i-th response is computed as:

$$A_{i,t} = \frac{r_i - mean(\{r_1, r_2, ..., r_G\})}{std(\{r1, r_2, ..., r_G\})} \tag{1}$$

The model is then optimized by maximizing the following objective:

$$\mathcal{J}_{GRPO}(\theta)$$
$$= \mathbb{E}_{q\sim P(Q)} \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}A_{i,t},\right.\right.$$
$$\left.\left.\text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}, 1-\varepsilon, 1+\varepsilon\right)A_{i,t}\right] - \beta\mathbb{D}_{KL}\left[\pi_\theta||\pi_{ref}\right]\right\} \tag{2}$$

where ε and β are hyper-parameters.

In our CR^3 method, the critical reward model adopts rule-based reward functions, which differ from conventional reward models in standard RL frameworks. The adopted reward functions include:

- **Accuracy reward** r_{acc}: this reward function checks whether the predicted output exactly matches the ground-truth answer. If they match identically, it returns a reward score of 1; otherwise, the score is 0. This straightforward reward scheme can mitigate the issue of reward hacking in reinforcement learning.
- **Format reward** r_{format}: the format reward verifies whether the model's output adheres to a required format. The adopted format prompt instructs the model to: "first output the thinking process in <think> </think> tags and then output the final answer in <answer> </answer> tags". The reward score is 1 only when the output strictly follows this format; otherwise, the score is 0. This function enforces explicit reasoning generation while avoiding content-specific biases.

The final rule-based reward function combines the accuracy reward r_{acc} with the format reward r_{format} as follows:

$$r = r_{acc} + \lambda r_{format} \tag{3}$$

where λ represents the format reward weight, controlling the relative importance between the accuracy reward and the format reward. The rule-based rewards provides accurate and reliable feedback for policy model, thereby minimizing the impact of noisy or ambiguous signals during training.

3.2 Data Selection via Multimodal Filtering

Effective model training relies fundamentally on high-quality training data. However, existing composition-aware image-text datasets (e.g., TripletData [1] (Patel et al. 2024), GMN (Sahin et al. 2024)) frequently contain noisy or trivial samples where multimodal match depends on simple entity detection rather than complex compositional relationships. Such samples can undermine both model effectiveness in capturing compositional information and learning stability in rule-based reinforcement learning. To address this, we propose a multimodal collaborative filtering strategy to distill TripletData into a high-quality dataset tailored for advanced compositional reasoning. We first randomly sample 185,000 instances as shown in Figure 2 from the original TripletData dataset. Our filtering process then removes samples where the positive and hard-negative pairs are too dissimilar in either modality, thus ensuring the remaining examples require genuine compositional reasoning.

[1]TripletData is released at https://huggingface.co/datasets/TripletCLIP/TripletCLIP-High-Quality

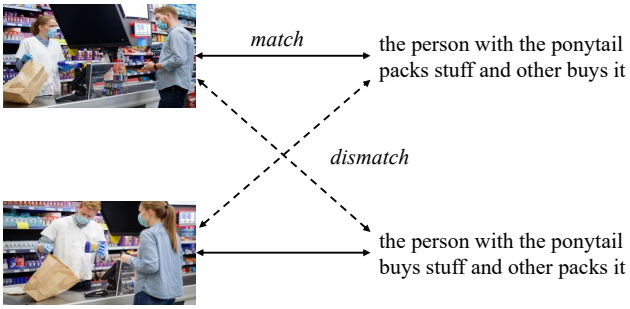


Figure 2: Instances of TripletData. Each sample contains two matched image-text pairs (marked with solid lines). The two image-text pairs differ in their compositional information. Consequently, the mismatched pair, indicated by dashed lines, serves as a compositionally-aware hard negative for the matched pair.

Textual Filtering: We employ SBERT (Reimers and Gurevych 2019) to estimate the semantic similarity between two captions in each data sample. Inspired by the high textual similarity in the human-curated Winoground benchmark (average score: 0.97) (Thrush et al. 2022), we set a similarity threshold of 0.7. Samples with scores below this threshold are discarded, ensuring the retrained samples require challenging textual compositional reasoning.

Visual Filtering: Similarly, we use DINOv2 (Oquab et al. 2023) to measure the similarity between two images in each data sample. Guided by the visual similarity distribution in Winoground, we apply a threshold of 0.75. This step filters out pairs where the images are visually distinct, forcing the model to focus on fine-grained spatial and relational details.

TripletData	Text Score	Img Score	Grp Score
before	67.9	51.3	42.3
after	54.3	46.7	33.9

Table 1: Winoground-style evaluation of Qwen2.5-VL-7B on TripletData before and after our filtering. The text, image, and group scores assess its textual, visual, and multimodal reasoning capabilities, respectively, with lower scores indicating greater difficulty.

This stringent filtering process discards approximately 90% of the initial samples, yielding a condensed, high-quality dataset of 18,900 instances. As shown in Table 1, the compositional difficulty of the dataset, evaluated using a SoTA MLLM, increases significantly after filtering. This curated dataset forms the bedrock of our training framework.

3.3 Compositional Reasoning Tasks for MLLMs

Building on our curated dataset, we design three distinct yet complementary tasks to holistically enhance MLLMs' compositional reasoning capabilities. Each task is formulated with simple rules, making them particularly suitable for reinforcement learning with verifiable rewards.

Task	Prompt
text-guided visual compositional reasoning	First image: $\{image_1\}$ Second image: $\{image_2\}$ Which image best matches the caption below? Caption: $\{Caption_1\}$ Output the final answer with First or Second.
visual-guided textual compositional reasoning	$\{image\}$ Which caption best describes the given image? A.$\{Caption_1\}$ B.$\{Caption_2\}$ Output the final answer with the option's letter A or B.
compositional image-text matching	$\{image_1\}$ Does the below caption precisely describe the given image? Caption: $\{Caption_1\}$ Output the final answer with Yes or No.

Table 2: The prompts used for different types of compositional reasoning training tasks. Note that the format prompt to enclose the reasoning processes within <think> and </think> tags is omitted due to space limitations.

- **Text-Guided Visual Compositional Reasoning** (TG-VCR): In this text-to-image alignment task, the MLLM is given a caption and must select the corresponding image from two options: the correct one and a hard negative with compositional difference. This task specifically trains the model to perform compositional reasoning about visual information based on textual semantic guidance.
- **Visual-Guided Textual Compositional Reasoning** (VG-TCR): As the inverse counterpart to TG-VCR, this image-to-text alignment task requires selecting the correct caption from a pair of textual options given an input image. This task complements the above one by learning bidirectional compositional reasoning (text-to-image and image-to-text), critical for robust multimodal models.
- **Compositional Image-Text Matching** (CITM): In this binary classification task, the model must determine whether an image-text pair constitutes a precise match. By exclusively employing hard negatives as negative samples, the task requires direct compositional verification rather than comparative analysis, thereby facilitating more profound alignment understanding.

Original samples from the curated dataset are transformed into three compositional reasoning tasks by applying corresponding prompt templates, as detailed in Table 2. To mitigate positional bias in the TG-VCR and VG-TCR tasks, the order of candidate answers is randomized. For the CITM task, a balanced 1:1 ratio of positive to negative samples is maintained.

3.4 Model-Adaption Dynamic Mixing Strategy

A fundamental challenge in enhancing MLLMs' compositional reasoning abilities through reinforcement learning lies in their heterogeneous performance across different tasks, which renders the data mixing strategy crucial for effective training. To overcome this, we propose a model-adaptive dynamic task mixing strategy that automatically adjusts the training data distribution based on the model's evolving performance, thus optimizing the learning trajectory.

During training, we evaluate the model every 200 steps on a held-out validation set (1000 samples from our curated data, formatted for all three tasks). The resulting task-specific performance scores are then used to dynamically adjust the data sampling proportion for the subsequent training stage. Based on the principle that the task with lower performance require a greater need for data exposure, we formulate the sampling proportion p_i for each task i on model m as:

$$p_i^m = \frac{\prod_{j \neq i}(s_j^m + \alpha)}{\sum_{k=1}^{3} \prod_{l \neq k}(s_l^m + \alpha)} \tag{4}$$

where α is a smoothing term and s_i^m is the performance score (accuracy) for task i in model m. Note that the proportions satisfy $\sum_{i=1}^{3} p_i^m = 1$, ensuring a valid probability distribution. This self-regulating mechanism dynamically allocates more resources to underperforming tasks, thereby improving training stability and efficiency while eliminating manual tuning of data mixing ratios.

4 Experiments

4.1 Implementation Details

The CR^3 method adopts SoTA MLLMs, Qwen2.5-VL-3B/7B-Instruct(Bai et al. 2025) and InternVL3-2/8B(Zhu et al. 2025), as baselines. For the GRPO algorithm, we configure the total batch size to 16, the sampling number for each question to 8, and the maximum generation length to 1024 to maintain sufficient reasoning capacity during training. In our experiments, the KL divergence penalty is disabled (β=0) to prevent suppression of deep reasoning capabilities, while the clipping hyperparameter ε is set to 0.2. The format reward scaling factor λ is fixed at 1.0 to achieve optimal performance [2] All baseline models are optimized using a learning rate of 1e-6 and a linear learning rate scheduler. Furthermore, we apply supervised fine-tuning (SFT) to the baselines using identical training data and hyperparameters as CR^3 method for a fair comparison.

4.2 Evaluation Benchmarks and Metrics

To comprehensively validate the effectiveness of our CR^3 method, we establish a dual-dimensional evaluation framework encompassing both in-domain and out-of-domain scenarios. For in-domain evaluation, we select three popular compositional reasoning benchmarks:

- **MMVP (Tong et al. 2024)**: its vision-question answering paradigm evaluates visual compositional reasoning

[2] A comprehensive ablation study on β is provided in the source code.

Method	MMVP	Winoground			Cola			Avg.
	Acc.	Text	Image	Group	Text	Image	Group	
Human	95.7	89.5	88.5	85.5	-	83.9	-	
Random	25.0	25.0	25.0	16.7	25.0	25.0	16.7	22.6
CLIP (ViT-B/32)	-	30.8	11.0	8.8	38.6	26.7	17.6	-
SigLIP 2 (ViT-so/14)	-	38.3	19.0	16.0	-	-	-	
GPT4O	70.7	62.0	58.3	44.3	76.2	58.1	50.5	60.0
Qwen2.5-VL-3B	26.0	61.8	10.8	9.0	75.2	1.4	1.4	26.6
+SFT	30.0	59.8	18.0	13.8	60.9	15.7	11.4	29.9
+CR^3	**44.7**	**66.8**	**32.8**	**27.0**	**78.6**	**33.3**	**29.1**	**44.6**
Qwen2.5-VL-7B	20.0	73.9	30.7	28.1	82.4	51.9	43.3	47.2
+SFT	44.0	73.1	34.4	32.7	79.1	50.9	41.4	50.8
+CR^3	**51.3**	**75.1**	**40.0**	**35.7**	**82.9**	**61.9**	**53.8**	**57.2**
InternVL3-2B	34.0	32.0	8.3	2.3	63.8	20.0	13.8	24.9
+SFT	37.3	37.6	19.3	9.3	65.7	22.4	15.2	29.5
+CR^3	**38.0**	**47.5**	**27.5**	**12.8**	**71.9**	**34.3**	**22.9**	**36.4**
InternVL3-8B	55.3	69.5	25.3	19.8	81.4	47.6	42.4	48.8
+SFT	56.0	70.0	27.3	22.3	81.4	50.9	43.8	50.2
+CR^3	**59.3**	**72.0**	**45.0**	**36.8**	**84.3**	**57.6**	**51.9**	**58.1**

Table 3: Zero-shot performance on in-domain compositional reasoning benchmarks. Best performance among the vanilla baseline, the SFT method and our CR^3 method are highlighted in **bold**.

by requiring models to answer paired questions associated with two compositional different images. A sample is scored only if both answers are correct.

- **Winoground (Thrush et al. 2022) & Cola (Ray et al. 2023)**: they employ an image-text matching framework, where each sample contains two matched image-text pairs forming challenging hard negative pairs (as shown in Fig. 2. Winoground introduces three metrics: text score (image-to-text retrieval accuracy), image score (text-to-image retrieval accuracy) and group score (correct retrieval in both directions). These metrics focus on textual, visual and multimodal compositional reasoning, respectively.

For out-of-domain evaluation, we adopts multiple popular benchmarks (including MMB (Liu et al. 2024), MME (Fu et al. 2023), MMMU (Yue et al. 2024), HallusionBench (Guan et al. 2024), MMStar (Chen et al. 2024b)) and fine-grained multimodal tasks, like OCRBench, visual spatial reasoning (VSR) and object counting (TallyQA).

4.3 Results on In-Domain Benchmarks

Tables 3 presents the performance of baseline models, SFT, and our proposed CR^3 on the three compositional reasoning benchmarks. Obviously, our CR^3 method demonstrates significant performance improvement across different architectures and scales of MLLMs. In specific, CR^3 achieves average absolute gains of 18.0 and 11.5 compared to Qwen2.5-VL-3B and InternVL3-2B baselines, respectively. It also boosts the average compositional performance of Qwen2.5-VL-7B and InternVL3-8B by 10 absolute points, largely bridging the performance gap to the advanced GPT-4o. Moreover, CR^3 achieves an absolute average improvement of over 5 points compared to SFT-based method, highlighting the effectiveness of rule-based reinforcement learning in enhancing compositional reasoning of MLLMs.

Compared to baselines, both SFT and CR^3 approaches achieve significantly visual compositional reasoning gains and relatively modest performance gains on the text scores of Winoground and Cola. This is primarily due to the strong textual understanding capabilities inherent in LLM-centric MLLMs, which makes enhancing visual compositional reasoning a more rewarding optimization direction during training. As analyzed in Section 5.1, CR^3's dynamic mixing strategy can partially mitigate this issue, but it remains an open challenge requiring further research.

4.4 Results on Out-of-Domain Benchmarks

To assess the generalization of our method beyond compositional tasks, we evaluated CR^3 on a suite of out-of-domain multimodal benchmarks. The results, presented in Table 4, reveal two critical findings. First, CR^3 consistently surpasses all baselines, delivering measurable improvements in general multimodal understanding and fine-grained vision tasks such as OCR, VSR, and TallyQA. Second, this stands in stark contrast to standard SFT approaches, which exhibit performance degradation across these diverse tasks. This divergence underscores the fundamental advantage of our rule-based reinforcement learning approach. Unlike SFT, which tends to fit to specific data distributions, CR^3 enhances the model's intrinsic capabilities through rule-guided self-exploration. This process fosters robust generalization rather than simple pattern matching.

Method	MMB	MME	MMMU	Hallu.	MMStar	OCR.	VSR	TallyQA
Qwen2.5-VL-3B	**80.0**	2169	47.1	43.4	54.1	82.7	76.2	81.9 / 72.5
+SFT	79.6	2164	46.7	43.0	**57.3**	82.2	71.1	82.0 / 72.1
+CR^3	79.4	**2201**	**47.3**	**43.4**	55.5	**83.3**	**76.7**	**82.4 / 73.0**
Qwen2.5-VL-7B	83.0	2302	46.7	47.3	61.8	88.4	73.8	84.9 / **74.4**
+SFT	83.5	2306	46.0	48.4	61.3	88.2	72.9	84.9 / 74.4
+CR^3	**85.2**	**2346**	**52.0**	**49.5**	**64.8**	**88.7**	**80.9**	**85.0** / 74.2
InternVL3-2B	81.4	2183	44.7	42.1	61.3	83.3	71.3	83.9 / 71.1
+SFT	80.7	2144	42.7	43.7	58.8	83.0	70.1	83.1 / 72.4
+CR^3	**81.8**	**2193**	**47.3**	**44.0**	**61.5**	**84.0**	**71.8**	**84.2 / 73.1**
InternVL3-8B	85.9	2411	57.3	49.3	68.6	88.1	80.2	85.5 / 71.2
+SFT	86.3	2403	56.0	51.4	**68.8**	87.6	80.5	85.3 / 74.5
+CR^3	**86.4**	**2428**	**58.3**	**51.5**	68.7	**88.2**	**82.3**	**85.5 / 76.0**

Table 4: Performance on out-of-domain general multimodal tasks. "Hallu." refers to the HallusionBench, "OCR." refers to the OCRBench. The results of TallyQA on "simple" and "complex" questions are presented separately.

Furthermore, the strong performance of CR^3 on out-of-domain tasks underscores compositional reasoning as a foundational competence for multimodal understanding. Enhancing compositional reasoning not only facilitates deeper comprehension of image-text semantic structures but also overcomes key limitations in fine-grained vision-language tasks. Remarkably, CR^3 achieves these advancements with only 18,000 training samples, a data size that is orders-of-magnitude smaller than that required by the MLLM baselines. This dramatic data efficiency, coupled with its emergent reasoning capabilities, validates the necessity of prioritizing compositional reasoning in multimodal learning frameworks.

5 In-Depth Analysis

5.1 The Influence of Compositional Tasks

To quantify the impact of the three compositional training tasks TG-VCR, VG-TCR and CITM, which are corresponding to the ability of visual compositional reasoning, textual compositional reasoning, and multimodal compositional alignment respectively, we conduct a comprehensive ablation study on the Qwen2.5-VL-3B baseline.

As shown in Table 5, the TG-VCR task significantly enhances performance on visual composition challenges, while VG-TCR yields substantial gains in textual compositional reasoning. Meanwhile, the CITM task leads to superior results on the MMVP benchmark, highlighting its strength in aligning vision and language modalities for question answering. These targeted improvements validate our methodological design, showing that the task suite collectively addresses critical aspects of compositional reasoning in MLLMs. Most notably, the combined implementation reveals significant synergistic effects - the full-task model outperforms all single-task variants across every evaluation metric, proving that our components are not only individually effective but also mutually reinforcing.

Training Tasks	**MMVP**	**Winoground**		
	Acc.	Text	Image	Group
Qwen2.5-VL-3B	26.0	61.8	10.8	9.0
+ VG-TCR only	32.7	**65.8**	16.0	13.5
+ TG-VCR only	36.0	63.0	**27.5**	**20.5**
+ CITM only	**39.3**	62.3	15.0	12.5
+ uniform mixing	38.0	63.5	25.0	18.3
+ fixed-ratio mixing	41.3	63.8	27.5	20.5
+ dynamic mixing	**44.7**	**66.8**	**32.8**	**27.0**

Table 5: Results on Qwen2.5-VL-3B with different training tasks and mixing strategies.

5.2 Effectiveness of Task Dynamic Mixing

We validate our dynamic task-mixing strategy through comprehensive comparisons with models trained using static uniform sampling and fixed-ratio sampling policies. As shown in Table 5, our approach consistently outperforms these baselines, achieving superior performance on all evaluation metrics. We attribute this improvement to the reinforcement learning-guided scheduler, which establishes an implicit curriculum by progressively adjusting the task mixture towards greater difficulty. The effectiveness of this strategy confirms that an adaptive training regimen with progressively increasing difficulty significantly enhances the model's compositional reasoning capabilities.

5.3 The Impact of Format Reward

Structured output generation is critical, yet the potential of format rewards remains underexplored in current research. In this study, we investigate the impact of varying the format reward scaling factor λ within the CR^3 method on compositional reasoning performance. Comprehensive results are summarized in Table 6. Remarkably, optimal performance on both the Winoground and Cola benchmarks is attained at

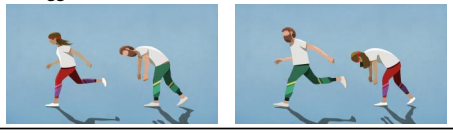
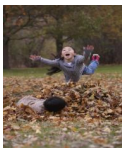
Input Sample	Thinking & Answer
Which image best match the below caption? Caption: the person with green legs is running quite slowly and the red legged one runs faster 	<think> The caption describes two people, one with green legs and the other with red legs. The person with green legs is running slowly, while the person with red legs is running faster. This description matches the first illustration where the person on the right (green legs) is bending over, indicating they are running slowly, and the person on the left (red legs) is running forward. </think> <answer>First</answer>
Which caption best describe the given image? A.the orange lollipop is surprised and the red lollipop is sad B.the orange lollipop is sad and the red lollipop is surprised 	<think> The question requires a choice between two captions describing lollipops with facial features. Option A describes the orange lollipop as surprised and the red lollipop as sad, which is a common interpretation of the images. Option B reverses this, describing the orange lollipop as sad and the red lollipop as surprised. The standard convention is to describe the leftmost figure first. </think> <answer>B</answer>
Which image best match the below caption? Caption: the leaves fall on people  	<think> The caption "the leaves fall on people" suggests that leaves are falling onto individuals. In the first image, leaves are falling from above towards two people, indicating that the leaves are falling on them. The second image shows children playing in leaves, but there is no indication of leaves falling specifically on them. Therefore, the first image better matches the caption. </think> <answer>First</answer>

Figure 3: Qualitative results on Winoground generated by CR^3.

λ	Winoground			Cola		
	Text	Image	Group	Text	Image	Group
0.0	62.8	26.8	19.5	76.2	21.4	16.7
0.3	63.5	32.5	24.0	78.1	28.1	24.8
0.5	64.5	31.8	23.0	70.0	31.4	23.8
0.8	63.3	30.5	23.8	74.3	28.1	22.9
1.0	**66.8**	**32.8**	**27.0**	**78.6**	**33.3**	**29.1**
1.3	65.0	32.3	24.5	66.2	30.0	24.8

Table 6: Results of CR^3 on Winoground and Cola benchmarks with different format reward scaling factor.

$\lambda = 1.0$. This result is counterintuitive, as one might anticipate that diminishing the format reward's influence would enable the model to prioritize task accuracy, thus enhancing compositional reasoning. However, our findings reveal that even an increase in λ to 1.3 leads to suboptimal performance compared to $\lambda = 1.0$.

5.4 Case Analysis

To qualitatively assess the effectiveness of CR^3, we showcase representative examples from the Winoground benchmark in Figure 3, presenting the model's reasoning processes alongside the corresponding answers. These cases show that for each multi-modal input, CR^3 initially conducts a thorough analysis of the inquiry, explicitly identifying and reasoning about visually-grounded elements during this process. This structured reasoning process allows the model to better comprehend the interrelationships among object states, thereby mitigating the characteristic errors illustrated in Figure 1. The reasoning traces of the model demonstrate its proficiency in accurately interpreting compositional information from both visual and textual modalities, effectively aligning cross-modal representations to generate accurate responses.

6 Conclusions

In this work, we introduced CR^3, a novel framework that pioneers rule-based reinforcement learning to enhance compositional reasoning in MLLMs. By integrating rigorous data curation with a model-adaptive dynamic mixing strategy, CR^3 systematically optimizes MLLMs' ability to reason about compositional information. Extensive experiments demonstrate that CR^3 achieves consistent improvements (19%+) across diverse benchmarks, significantly outperforming SFT methods while exhibiting superior generalization. The ablation studies further validated the effectiveness of our data curation and dynamic training strategy. These findings establish CR^3 as a promising paradigm for compositional reasoning enhancement, suggesting that rule-based RL offers superior data efficiency and generalization compared to conventional SFT approaches. To promote reproducibility and community progress, we release our compositionally-aware visual instruction-following dataset. Future work could explore extending this rule-based RL framework to hierarchical reasoning and multimodal knowledge transfer, which may facilitate interpretable and robust AI systems.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62176074).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Chen, B.; Xu, Z.; Kirmani, S.; Ichter, B.; Sadigh, D.; Guibas, L.; and Xia, F. 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14455–14465.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; et al. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Deng, Y.; Bansal, H.; Yin, F.; Peng, N.; Wang, W.; and Chang, K.-W. 2025. OpenVLThinker: An Early Exploration to Complex Vision-Language Reasoning via Iterative Self-Improvement. *arXiv preprint arXiv:2503.17352*.

Doveh, S.; Arbelle, A.; Harary, S.; Herzig, R.; Kim, D.; Cascante-Bonilla, P.; Alfassy, A.; Panda, R.; Giryes, R.; Feris, R.; et al. 2023. Dense and aligned captions (dac) promote compositional reasoning in vl models. *Advances in Neural Information Processing Systems*, 36: 76137–76150.

Feng, K.; Gong, K.; Li, B.; Guo, Z.; Wang, Y.; Peng, T.; Wang, B.; and Yue, X. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.

Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14375–14385.

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

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

Janssen, T. M.; and Partee, B. H. 1997. Compositionality. In *Handbook of logic and language*, 417–473. Elsevier.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, 216–233. Springer.

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

Ma, Z.; Hong, J.; Gul, M. O.; Gandhi, M.; Gao, I.; and Krishna, R. 2023. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10910–10921.

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

Ni, R.; Xiao, D.; Meng, Q.; Li, X.; Zheng, S.; and Liang, H. 2025. Benchmarking and understanding compositional relational reasoning of llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19703–19711.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Patel, M.; Kusumba, A.; Cheng, S.; Kim, C.; Gokhale, T.; Baral, C.; and Yang, Y. 2024. TripletCLIP: Improving Compositional Reasoning of CLIP via Synthetic Vision-Language Negatives. *Advances in neural information processing systems*.

Peng, W.; Xie, S.; You, Z.; Lan, S.; and Wu, Z. 2024. Synthesize Diagnose and Optimize: Towards Fine-Grained Vision-Language Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13279–13288.

Peng, Y.; Wang, X.; Wei, Y.; Pei, J.; Qiu, W.; Jian, A.; Hao, Y.; Pan, J.; Xie, T.; Ge, L.; et al. 2025. Skywork R1V: Pioneering Multimodal Reasoning with Chain-of-Thought. *arXiv preprint arXiv:2504.05599*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ray, A.; Radenovic, F.; Dubey, A.; Plummer, B.; Krishna, R.; and Saenko, K. 2023. Cola: A benchmark for compositional text-to-image retrieval. *Advances in Neural Information Processing Systems*, 36: 46433–46445.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sahin, U.; Li, H.; Khan, Q.; Cremers, D.; and Tresp, V. 2024. Enhancing multimodal compositional reasoning of visual language models with generative negative mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5563–5573.

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

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. FLAVA: A Foundational Language And Vision Alignment Model. In *CVPR*.

Stone, A.; Soltau, H.; Geirhos, R.; Yi, X.; Xia, Y.; Cao, B.; Chen, K.; Ogale, A.; and Shlens, J. 2025. Learning visual composition through improved semantic guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3740–3750.

Team, K.; Du, A.; Gao, B.; Xing, B.; Jiang, C.; Chen, C.; Li, C.; Xiao, C.; Du, C.; Liao, C.; et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Thrush, T.; Jiang, R.; Bartolo, M.; Singh, A.; Williams, A.; Kiela, D.; and Ross, C. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5238–5248.

Tong, S.; Liu, Z.; Zhai, Y.; Ma, Y.; LeCun, Y.; and Xie, S. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9568–9578.

Xie, Z.; Lin, M.; Liu, Z.; Wu, P.; Yan, S.; and Miao, C. 2025. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*.

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

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.

Yuksekgonul, M.; Bianchi, F.; Kalluri, P.; Jurafsky, D.; and Zou, J. 2023. When and why Vision-Language Models behave like Bags-of-Words, and what to do about it? In *International Conference on Learning Representations*.

Zeng, Y.; Zhang, X.; and Li, H. 2022. Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts. In *International Conference on Machine Learning*, 25994–26009. PMLR.

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

Zhou, H.; Li, X.; Wang, R.; Cheng, M.; Zhou, T.; and Hsieh, C.-J. 2025. R1-Zero's" Aha Moment" in Visual Reasoning on a 2B Non-SFT Model. *arXiv preprint arXiv:2503.05132*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.