Co-Reinforcement Learning for Unified Multimodal Understanding and Generation

Jingjing Jiang[1,2] **Chongjie Si**[1] **Jun Luo**[2] **Hanwang Zhang**[2] **Chao Ma**[1,*]

[1] Shanghai Jiao Tong University [2] Nanyang Technological University

jingjingjiang2017@gmail.com, {chongjiesi,chaoma}@sjtu.edu.cn
{junluo,hanwangzhang}@ntu.edu.sg

Abstract

This paper presents a pioneering exploration of reinforcement learning (RL) via group relative policy optimization for unified multimodal large language models (ULMs), aimed at simultaneously reinforcing generation and understanding capabilities. Through systematic pilot studies, we uncover the significant potential of ULMs to enable the synergistic co-evolution of dual capabilities within a shared policy optimization framework. Building on this insight, we introduce **CoRL**, a **Co**-**R**einforcement **L**earning framework comprising a unified RL stage for joint optimization and a refined RL stage for task-specific enhancement. With the proposed CoRL, our resulting model, **ULM-R1**, achieves average improvements of 7% on three text-to-image generation datasets and 23% on nine multimodal understanding benchmarks. These results demonstrate the effectiveness of CoRL and highlight the substantial benefits of reinforcement learning in facilitating cross-task synergy and optimization for ULMs. Code is available at https://github.com/mm-vl/ULM-R1.

1 Introduction

As large foundation models (LFMs) continue to advance in their general capabilities and breadth of knowledge, post-training [30, 46, 70, 73, 108] has emerged as a critical paradigm for further refining pretrained LFMs toward specialized applications, thereby facilitating task adaptation and human-aligned behaviors. Recently, reinforcement learning (RL)-based approaches [51, 52, 59, 64, 69, 72, 95] have exhibited considerable promise due to their data efficiency and strong alignment abilities. A notable exemplar is DeepSeek-R1 [19], which demonstrates that RL with verifiable rewards and the group relative policy optimization (GRPO) algorithm constitutes a practical and stable strategy that sidesteps explicit preference modeling [77] and reward model learning [83]. This promising paradigm indicates *the significant potential of LFMs to acquire advanced capabilities and generalize effectively without dependence on large-scale, high-quality supervised data*.

In the multimodal AI research community, the prevailing implementation [9, 24, 40, 41, 53, 69, 100, 101] of the GRPO algorithm centers on crafting diverse rule-based reward mechanisms to incentivize long-chain reasoning capabilities of multimodal large language models (MLLMs). These initiatives primarily target multimodal understanding, with a particular focus on visual and mathematical reasoning tasks. Conversely, its application to visual generation remains surprisingly limited, with only pioneering explorations [25, 79] suggesting its feasibility. More importantly, extending GRPO to unified MLLMs (ULMs) [8, 39, 88, 90] capable of concurrently performing visual understanding and generation tasks remains considerably under-explored. Intuitively, ULMs could significantly

*Corresponding author.

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

benefit from GRPO owing to their inherent advantages of *cross-task synergy* and *LLM sharing*, which enables ULMs to share reward signals across various tasks and effectively mitigate reward imbalance, particularly as GRPO operates by jointly ranking outputs within task-agnostic groups.

This paper aims to enhance the understanding and generation capabilities of ULMs without relying on supervised data. We begin with a set of pilot experiments to explore efficient reinforcement learning paradigms. Specifically, we systematically examine four rule-based training strategies: (*i*) separate RL for individual tasks, (*ii*) separate RL with weight merging, (*iii*) cycle RL alternating between tasks, and (*iv*) unified RL with joint optimization. Our explorations reveal two critical findings. ***First***, direct task-specific RL fails to achieve the anticipated improvements, particularly in visual generation, and even impairs other abilities. ***Second***, compared with alternative strategies, unified RL showcases comprehensive advantages across tasks. These results demonstrate the synergistic co-evolution of dual capabilities under a shared policy optimization paradigm.

In light of our preliminary findings, we propose **CoRL**, a co-reinforcement learning framework designed to synergistically improve the understanding and generation capabilities of ULMs. Specifically, CoRL follows a *Foundation-then-Specialization* paradigm and is implemented through a two-stage RL procedure: a unified RL stage for joint optimization of dual capabilities and a refined RL stage for task-specific enhancement. In the first stage, the policy ULM is optimized through a unified GRPO algorithm with diverse rewards on a carefully curated dataset spanning both understanding and generation tasks. To effectively guide policy optimization in visual generation, we introduce a *bidirectional cycle consistency reward* and a *text-image matching reward*, which together promote semantic consistency and faithfulness between synthesized images and their corresponding prompts. The designed rewards complement typical multimodal understanding rewards (*i.e.*, accuracy and format) within a unified group, enabling cross-task joint optimization. In the subsequent stage, we independently reinforce the policy's understanding and generation capabilities using respective rewards and tailored datasets for task-specific refinement.

Applying the two-stage CoRL procedure to the baseline ULM Janus-Pro [8] yields **ULM-R1**, a unified model with reinforced capabilities in both understanding and generation. To comprehensively assess its performance, we conduct extensive comparisons against state-of-the-art unified MLLMs and dedicated models across both three visual generation and nine multimodal understanding benchmarks. Notably, ULM-R1 achieves substantial gains over its baseline on complex mathematical and logical reasoning tasks, such as WeMath (+15.2) and LogicVista (+10.6). These results underscore the effectiveness of CoRL, providing compelling empirical evidence for the efficacy of RL in jointly advancing visual understanding and generation tasks.

We summarize our main contributions as follows:

- We establish that RL with verifiable rewards and GRPO constitutes a data-efficient paradigm for cross-task co-optimization and capability enhancement.
- We introduce a co-reinforcement learning framework, CoRL, to synergistically enhance the dual capabilities of ULMs using a unified-then-refined RL paradigm.
- We demonstrate the effectiveness of CoRL and the advantage of ULM-R1 through extensive qualitative and quantitative experiments across diverse benchmarks.

2 Related Work

Unified Multimodal Understanding and Generation. Recent advancements [8, 39, 67, 71, 76, 82, 86–88, 90, 91, 107] have witnessed increasing attention to jointly model multimodal understanding and visual generation within a unified model. Pioneering attempts [12, 17] predominantly rely on continuous diffusion models, integrating external diffusion decoders for image synthesis. Inspired by autoregressive next-token prediction, a growing line of research [8, 31, 39, 47, 57, 71, 82, 85–88, 111] encode visual inputs into discrete tokens and generate images in a *fully autoregressive* (F-AR) manner. Specifically, this approach employs a vector quantized (VQ) tokenizer [14, 93] to convert images into discrete tokens, analogous to text tokenization. To mitigate information loss in VQ discretization, another stream of work [4, 16, 23, 29, 48, 65, 74, 76, 90, 92] explores *autoregressive and diffusion* (AR-Diff) hybrid modeling approaches. Architecturally, these models typically comprise a vision autoencoder, a text tokenizer, and an LLM. Given the unified advantage of the F-AR model in generation manner, this work builds upon it to develop our co-reinforcement learning framework.

RL-based Post-Training for MLLMs. Post-training [108] aims to further enhance the performance of pretrained models for customized applications and user needs. Recently, RL [78, 84] has emerged as a powerful post-training technique, enabling models to learn from feedback and align with human values. RL in MLLMs can be broadly categorized into two paradigms: (1) RL from human/AI feedback (RLHF) [34, 54, 61, 68, 77, 81–83, 92, 95, 96, 99, 106, 109] and (2) RL with verifiable reward mechanisms [35, 40, 41, 69, 79, 100]. RLHF involves learning reward models from preference data before RL optimization, whereas the latter directly optimizes models using task-specific reward functions, bypassing explicit preference modeling. For example, DPO [59] is a notable implementation of RLHF and has been adopted by Emu3 [82] and HermesFlow [92] to narrow the performance gap between understanding and generation. In contrast, GRPO [64] exemplifies the second paradigm, simplifying reward formulation via group-wise relative advantage estimation. Our work also falls into this paradigm but diverges from prior work such as SimpleAR [79], which utilizes GRPO with external CLIP reward for autoregressive visual generation, and R1-like MLLMs [24, 41, 69, 100] that focus on incentivizing reasoning capabilities. First, our work demonstrates the significant potential of RL in co-optimizing understanding and generation, thereby broadening its applicability beyond reasoning. Moreover, we identify *semantic consistency rewards* and *a co-evolutionary reinforcement strategy* as crucial components in enhancing ULMs.

3 Methodology

3.1 Preliminary

Group relative policy optimization (GRPO) [64] is a value-free policy optimization algorithm with improved training stability and sample efficiency. Building upon PPO [62], GRPO introduces a group-wise relative advantage approach to bound policy updates while maintaining optimization flexibility. Let $\pi_{\boldsymbol{\theta}}$ denote a policy parameterized by $\boldsymbol{\theta}$. Formally, given an input content c, the algorithm first samples a group of G outputs $\{o_1, o_2, \ldots, o_G\}$ from the current policy $\pi_{\boldsymbol{\theta}_{\text{old}}}$. Each output is then evaluated using predefined, verifiable reward functions, yielding the reward set $\{r_1, r_2, \ldots, r_G\}$. These rewards are subsequently normalized to compute group-relative advantages as follows:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}. \tag{1}$$

After obtaining the advantage set $\{A_1, A_2, \ldots, A_G\}$ via group relative advantage estimation, the policy $\pi_{\boldsymbol{\theta}}$ is optimized by maximizing the following objective:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\{o_i\}_{i=1}^{G} \sim \pi_{\boldsymbol{\theta}_{\text{old}}}} \frac{1}{G} \sum_{i=1}^{G} \left[\frac{\pi_{\boldsymbol{\theta}}(o_i)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(o_i)} A_i - \beta \, \mathbb{D}_{\text{KL}} \left(\pi_{\boldsymbol{\theta}} \,\|\, \pi_{\text{ref}} \right) \right], \tag{2}$$

where $\mathbb{D}_{\text{KL}}$ denotes the KL-divergence used to constrain the deviation between $\pi_{\boldsymbol{\theta}}$ and its reference policy π_{ref}, and β is a regularization coefficient.

3.2 Pilot Exploration

Given the exceptional performance and data efficiency of DeepSeek-R1-Zero [19], we explore the potential of ULMs to enhance understanding and generation capabilities without dependence on task-specific supervised fine-tuning. To accomplish this, we curate a dataset[2] comprising 16K samples sourced from the COCO 2017 training split [38]. Each sample includes a real image, an associated caption as a textual prompt for visual generation, and a corresponding QA pair for the multimodal understanding task. We adopt CLIP Score [58] as the verifiable reward for image generation, along with a combination of formatting correctness and answer

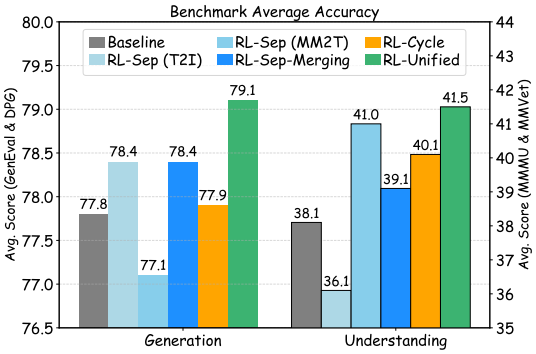

Figure 1: **Results of different RL paradigms.** Janus-Pro-1B [8] serves as the baseline.

[2] https://huggingface.co/datasets/mm-vl/x2x_rft_16k

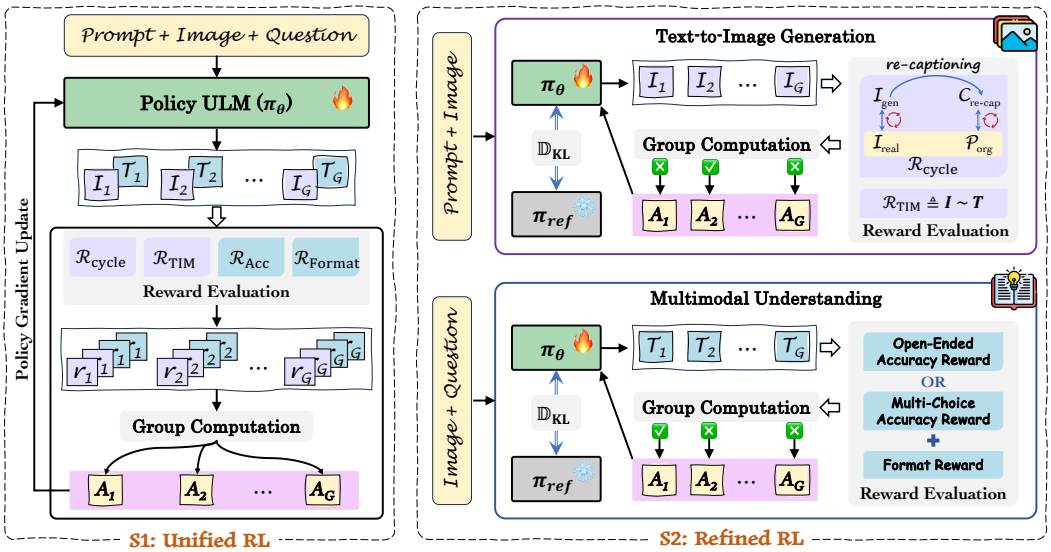

Figure 2: **Overview of CoRL**, a co-reinforcement learning framework to jointly improve the dual capabilities of ULMs. CoRL adopts a two-stage RL procedure, comprising a unified RL stage for joint optimization and a refined RL stage for task-specific enhancement.

accuracy as the reward for text generation. We investigate four distinct RL paradigms: (*i*) *separate RL*, where understanding and generation tasks are independently optimized with their respective reward mechanisms; (*ii*) *separate RL followed by weight merging*, where each task is separately optimized, and the resulting weights are subsequently merged using a Gaussian distribution-based merging strategy [66] to incorporate both abilities; (*iii*) *cycle RL*, which employs a scheduled alternation between the two tasks throughout the training process; and (*iv*) *unified RL*, in which both tasks are jointly optimized within a unified paradigm to promote the co-evolution of dual capabilities.

As presented in Figure 1, we observe that (1) direct task-specific RL fails to achieve the expected improvements for ULMs, particularly in the visual generation task, and may even impair performance on the other task; and (2) unified RL demonstrates substantial advantages over alternative paradigms. These findings indicate that the dual capabilities that co-evolve within a shared training framework contribute to enhanced cross-task synergy and knowledge transfer.

3.3 Co-Reinforcement Learning

3.3.1 Verifiable Reward for Multimodal modeling

In this section, we develop a suite of verifiable rewards for multimodal modeling, which provide clear and objective feedback to steer ULMs toward generating high-quality image and text outputs.

Bidirectional Cycle Consistency Reward in Text-to-Image Generation. To encourage ULMs to generate images that faithfully depict the concepts and entities described in the input prompt, we introduce a bidirectional cycle consistency reward $\mathcal{R}_{\text{cycle}}$, which measures the consistency between predictions and ground truth in both visual and textual spaces. For visual consistency, we adopt LPIPS [104] to assess the patch-level perceptual similarity between the real image $\mathcal{I}_{\text{real}}$ and the synthesized image $\mathcal{I}_{\text{gen}}$. Textual consistency is implemented in a re-captioning manner. Specifically, we first employ BLIP [32] to generate a caption $\mathcal{C}_{\text{re-cap}}$ for each synthesized image, and then compute the SPICE [1] score between $\mathcal{C}_{\text{re-cap}}$ and its original prompt $\mathcal{P}_{\text{org}}$ to measure semantic fidelity. The combined bidirectional cycle reward is defined as:

$$\mathcal{R}_{\text{cycle}} = 1 - \text{LPIPS}(\mathcal{I}_{\text{real}}, \mathcal{I}_{\text{gen}}) + \text{SPICE}(\mathcal{P}_{\text{org}}, \mathcal{C}_{\text{re-cap}}). \tag{3}$$

This bidirectional reward forms a closed feedback loop that promotes mutual consistency between texts and images, effectively penalizing hallucinated content and reinforcing prompt-aligned visual generation by simultaneously optimizing for both visual and textual consistency. Furthermore, $\mathcal{R}_{\text{cycle}}$ is normalized to the range [0, 1] before being combined to ensure that all rewards operate on comparable scales and to prevent any single component from dominating due to scale differences.

Text-Image Matching Reward. While CLIP Score [58] provides a holistic measure of text-image alignment, as shown in Sec. 3.2, it underperforms due to its limited capacity for assessing fine-grained semantics. To address this limitation, we instead propose a text-image matching reward $\mathcal{R}_{\text{TIM}}$, which leverages the ULM itself to evaluate cross-modal alignment at the token level. Given a textual representation $\boldsymbol{T} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_{L_t}\} \in \mathbb{R}^{L_t \times d}$ of the prompt and the corresponding visual representation $\boldsymbol{I} = \{\boldsymbol{i}_1, \boldsymbol{i}_2, \ldots, \boldsymbol{i}_{L_i}\} \in \mathbb{R}^{L_i \times d}$ of a generated image, the reward is computed as:

$$\mathcal{R}_{\text{TIM}} = \frac{1}{2}\left(\frac{1}{L_i}\sum_{j=1}^{L_i}\max_{k\in[1,L_t]}\cos(\boldsymbol{i}_j, \boldsymbol{t}_k) + \frac{1}{L_t}\sum_{k=1}^{L_t}\max_{j\in[1,L_i]}\cos(\boldsymbol{t}_k, \boldsymbol{i}_j)\right), \tag{4}$$

where L_t and L_i are the sequence lengths of the textual and visual tokens, d is the embedding dimension, and $\mathcal{R}_{\text{TIM}}$ is also be normalized to the range [0, 1]. This reward captures the fine-grained correspondence between textual concepts and visual elements through maximum cosine similarity, ensuring mutual alignment between visual tokens and their most relevant textual counterparts.

Accuracy Reward in Multimodal Question Answering. Accuracy rewards leverage task-specific metrics to directly evaluate the correctness of ULM predictions. We consider two accuracy rewards tailored to different question types: $\mathcal{R}_{\text{MCQ-Acc}}$ for multi-choice questions and $\mathcal{R}_{\text{OE-Acc}}$ for open-ended questions. These rewards follow a binary evaluation mechanism, assigning a value of 1 when the predicted answer (*i.e.*, the final answer parsed from within `<answer>` and `</answer>` tags) matches the ground truth and 0 otherwise.

Format Reward in Text Generation. To encourage ULMs to generate structured and interpretable textual responses, we adopt the format reward [19], which requires the model to enclose its thinking process inside `<think>` $\cdots$ `</think>`, and provide its final answer within `<answer>` and `</answer>` tags. The format reward $\mathcal{R}_{\text{Format}}$ returns 1 for strict compliance and 0 otherwise.

3.3.2 Unified Reinforcement Learning for Synergistic Multimodal Modeling

As illustrated in Figure 2, the policy ULM first undergoes unified reinforcement learning with diverse rewards across understanding and generation tasks. This unified process aims to jointly enhance its dual capabilities and establish a solid foundation for subsequent task-specific refinement.

Reward Function and Training Objective. To ensure diversity and complementarity in reward signals for unified multimodal modeling, we formulate a joint reward function as

$$\mathcal{R}_{\text{Uni-S1}} = \mathcal{R}_{\text{cycle}} + \mathcal{R}_{\text{TIM}} + \lambda \cdot (\mathcal{R}_{\text{Acc}} + \mathcal{R}_{\text{Format}}), \tag{5}$$

where λ is a coefficient that balances the two types of rewards. During training, given an input prompt and an image-question pair, the policy model $\pi_{\boldsymbol{\theta}_{\text{old}}}$ first generates G candidate responses, $o = \{(\mathcal{I}_1, \mathcal{T}_1), (\mathcal{I}_2, \mathcal{T}_2), \ldots, (\mathcal{I}_G, \mathcal{T}_G)\}$, each comprising a synthesized image $\mathcal{I}$ and a CoT-format solution $\mathcal{T}$. Concurrently, the joint reward function $\mathcal{R}_{\text{Uni-S1}}$ evaluates each candidate pair, yielding the reward set $r = \{r_1, r_2, \ldots, r_G\}$. These rewards are subsequently normalized according to Eq. (1) to compute the corresponding group-relative advantages $A = \{A_1, A_2, \ldots, A_G\}$. The new policy model $\pi_{\boldsymbol{\theta}}$ is then updated by maximizing the following GRPO-based objective:

$$\mathcal{L}_{\text{S1}} = \mathbb{E}_{\{o_i\}_{i=1}^{G}\sim\pi_{\boldsymbol{\theta}_{\text{old}}}}\frac{1}{G}\sum_{i=1}^{G}\frac{\pi_{\boldsymbol{\theta}}(o_i)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(o_i)}A_i\,, \text{ where } o_i = (\mathcal{I}_i, \mathcal{T}_i). \tag{6}$$

Notably, based on empirical findings from recent work [94], we omit the KL-divergence constraint during this stage to improve both optimization efficiency and generalization capability.

Training Data. To support unified RL for synergistic multimodal modeling, we curate a comprehensive dataset comprising 22K samples[3], which follows the data structure established in Sec. 3.2. Each sample includes *a real image*, *a prompt* for visual generation, and *a CoT-format QA pair* for multimodal understanding. This balanced data composition facilitates joint optimization of dual capabilities within a unified framework, while preserving the granularity of task-specific supervision.

[3] https://huggingface.co/datasets/mm-vl/x2x_rft_22k

Table 1: **Results on text-to-image generation benchmarks.** ♣ and ♣ denote models trained using DPO and GRPO strategies. The best performance in each category is highlighted in **bold**.

Model	Scale	Res.	Type	GenEval ↑					WISE ↑	DPG ↑
				Two Obj.	Counting	Position	Color Attri.	**Overall**	**Overall**	**Overall**
▼ *Generation Only*										
PixArt-α [5]	0.6B	512^2	Diff	0.50	0.44	0.08	0.07	0.48	**0.47**	71.11
SDv1.5 [60]	0.9B	512^2	Diff	0.38	0.35	0.04	0.06	0.43	0.32	63.18
SDv2.1 [60]	0.9B	512^2	Diff	0.51	0.44	0.07	0.17	0.50	0.32	68.09
SD3-Medium [15]	2B	512^2	Diff	**0.94**	**0.72**	**0.33**	**0.60**	**0.74**	0.42	**84.08**
SDXL [55]	2.6B	1024^2	Diff	0.74	0.39	0.15	0.23	0.55	0.43	74.65
DALL·E 3 [3]	-	1024^2	Diff	0.87	0.47	0.43	0.45	0.67	-	83.50
LlamaGen [67]	0.8B	256^2	F-AR	0.34	0.21	0.07	0.04	0.32	-	65.16
SimpleAR [79] ♣	1.5B	1024^2	F-AR	0.90	-	0.28	0.45	0.63	-	81.97
▼ *Unified Understanding and Generation*										
TokenFlow [57]	8B	256^2	F-AR	0.60	0.41	0.16	0.24	0.55	-	73.38
Emu3 [82]	8B	512^2	F-AR	-	-	-	-	0.66	0.39	80.60
Emu3-DPO [82] ♣	8B	512^2	F-AR	-	-	-	-	0.64	-	81.60
LWM [39]	7B	512^2	F-AR	0.41	0.46	0.09	0.15	0.47	-	-
Orthus [29]	7B	512^2	AR-Diff	-	-	-	-	0.58	0.27	-
Janus-Pro [8]	7B	384^2	F-AR	**0.89**	0.59	**0.79**	**0.88**	**0.80**	0.35	**84.19**
ILLUME+ [23]	3B	384^2	AR-Diff	0.88	0.62	0.42	0.53	0.72	-	-
D-DiT [37]	2B	512^2	Diff	0.80	0.54	0.32	0.50	0.65	-	-
Harmon [87]	1.5B	512^2	F-AR	0.86	0.66	0.74	0.48	0.76	**0.41**	-
show-o [90]	1.3B	512^2	AR-Diff	0.80	0.66	0.31	0.50	0.68	0.35	67.48
HermesFlow [92] ♣	1.3B	512^2	AR-Diff	0.84	0.66	0.32	0.52	0.69	-	70.22
Janus [85]	1.3B	384^2	F-AR	0.68	0.30	0.46	0.42	0.61	0.23	79.68
Janus-Pro [8]	1.5B	384^2	F-AR	0.82	0.51	0.65	0.56	0.73	0.26	82.63
ULM-R1 ♣	1.5B	384^2	F-AR	0.85	**0.71**	0.68	0.80	0.77	0.33	83.92

3.3.3 Refined Reinforcement Learning for Task-specific Enhancement

After completing unified RL, as shown in Figure 2, we apply a targeted learning strategy to further enhance the task-specific performance of the policy model. This second-stage optimization leverages task-specific rewards and tailored datasets for individual tasks.

Reward Function and Training Objective. For text-to-image generation, the reward is defined as $\mathcal{R}_{\text{T2I-S2}} = \mathcal{R}_{\text{cycle}} + \mathcal{R}_{\text{TIM}}$. For multimodal understanding, we define two distinct reward formulations: (1) $\mathcal{R}_{\text{MCQ-S2}} = \mathcal{R}_{\text{MCQ-Acc}} + \mathcal{R}_{\text{Format}}$ for multiple-choice questions, and (2) $\mathcal{R}_{\text{OE-S2}} = \mathcal{R}_{\text{OE-Acc}} + \mathcal{R}_{\text{Format}}$ for open-ended questions. The training objective in this stage adheres to the standard GRPO formulation in Eq. (2), with the appropriate task-specific reward ($\mathcal{R}_{\text{T2I-S2}}$, $\mathcal{R}_{\text{MCQ-S2}}$, or $\mathcal{R}_{\text{OE-S2}}$) replacing A_i depending on the task. To ensure stable optimization, we reintroduce the KL-divergence constraint at this stage to limit policy deviation from the reference distribution.

Training Data. For text-to-image generation, we continue training on the curated dataset introduced in Sec. 3.2. For multimodal understanding, we utilize two specialized datasets: `mcot_r1_mcq`[4] for multiple-choice questions and `mcot_r1_vqa`[5] for open-ended questions. These task-specific datasets enable the model to develop more refined and robust capabilities within each task domain.

4 Experiment

4.1 Experimental Setups

Evaluation Benchmarks. We evaluate visual generation capabilities on the GenEval [18], WISE [50], and DPG-Bench [22] benchmarks. GenEval employs an object-centric evaluation protocol to assess compositional and attribute-level alignment, while DPG-Bench adopts a VQA-based setting to evaluate dense prompt-following and semantic fidelity. WISE provides a holistic evaluation of models' world knowledge, considering consistency, realism, and aesthetics. We also evaluate multimodal understanding capabilities across diverse benchmarks. Specifically, MMStar [6], MMMU [98], and

[4] https://huggingface.co/datasets/mm-vl/mcot_r1_mcq_66k
[5] https://huggingface.co/datasets/mm-vl/mcot_r1_vqa_66k

Table 2: **Results on multimodal understanding benchmarks.** The best performance within each category is highlighted in **bold**. † denotes results obtained from our evaluation.

Model	LLM	Multi-Choice (MC) ↑			Open-Ended (OE) ↑			MC&OE Mixed ↑		
		MMMU	MMStar	Math^{We}	MMVet	POPE	Logic^{VT}	Math^{VT}	Math^{VS}	Math^{Vis}
▼ *Understanding Only*										
SmolVLM [49]	SmolLM2-1.7B	38.8	41.7	9.1	33.8	85.5	28.0	43.6	12.6	12.8
SAIL-VL [11]	Qwen2.5-1.5B	44.1	56.5	14.6	44.2	88.1	30.4	62.8	17.4	17.3
Ovis2 [45]	Qwen2.5-1.5B	45.6	56.7	9.9	58.3	87.8	34.7	**64.1**	29.4	17.7
InternVL3 [110]	Qwen2.5-1.5B	48.7	**61.1**	**22.9**	**67.0**	**90.1**	34.7	57.6	24.5	20.2
Qwen2.5-VL [2]	Qwen2.5-3B	**51.2**	56.3	**22.9**	60.0	85.9	**40.3**	61.2	31.2	21.9
LMM-R1 [53]	Qwen2.5-3B	-	58.0	-	-	-	-	63.2	**41.6**	**26.4**
▼ *Unified Understanding and Generation*										
ILLUME+ [23]	Qwen2.5-3B	**44.3**	-	-	40.3	87.6	-	-	-	-
Harmon [87]	Qwen2.5-1.5B	38.9	-	-	-	87.6	-	-	-	-
VILA-U [88]	LLaMA-2-7B	-	-	-	33.5	85.8	-	-	-	-
Orthus [29]	Chameleon-7B	28.2	-	-	-	79.6	-	-	-	-
UniToken [27]	Chameleon-7B	32.8	46.1	-	-	-	-	38.5	-	
SGen-VL [31]	InternLM2-1.8B	34.2	-	-	34.5	85.3	-	**42.7**	-	
Show-o [90]	Phi-1.3B	26.7	-	-	-	80.0	-	-	-	-
HermesFlow [92]	Phi-1.3B	28.3	-	-	-	81.4	-	-	-	-
Janus-Pro [8]	DeepSeek-LLM-7B	41.0	46.5	9.7	**50.0**	87.4	28.0	42.5	15.9	14.7
Janus [85]	DeepSeek-LLM-1.3B	30.5	37.6	3.4†	34.3	87.0	23.9†	33.7	14.9†	13.4†
Janus-Pro [8]	DeepSeek-LLM-1.5B	36.3	43.1†	5.9†	39.8	86.2	23.9†	37.3†	13.5†	13.4†
ULM-R1	DeepSeek-LLM-1.5B	42.3	**47.6**	**21.1**	43.9	**88.9**	**34.5**	42.5	**25.4**	**22.0**

WeMath (Math^{We}) [56] are used for multi-choice evaluation, while MMVet [97], POPE [36], and LogicVista (Logic^{VT}) [89] are used for open-ended evaluation. In addition, we employ MathVista (Math^{VT}) [43], MathVerse-Vision (Math^{VS}) [102], and MathVision (Math^{Vis}) [80] to assess complex mathematical reasoning capabilities, covering both multi-choice and open-ended QA formats. On these benchmarks, we compute accuracy using the toolkit VLMEvalKit [13].

Implementation Details. We develop ULM-R1 using Janus-Pro-1B [8] as the baseline ULM for unified multimodal understanding and generation. To ensure reproducibility and scalability, our RL training is built upon the trl [75] framework. In the unified RL stage, we employ the AdamW optimizer with an initial learning rate of 4e-6 and a batch size of 16. We sample 8 responses for both understanding and generation tasks, and set the reward balancing factor in Eq. (5) to 0.8. In the refined RL stage, we sample 16 responses for both multimodal understanding and text-to-image generation tasks. Additionally, we reduce the learning rate to 1e-6 to facilitate fine-grained optimization. All training is conducted on 8 NVIDIA H20 (96G) GPUs. During inference, greedy decoding is used for text generation in multimodal understanding tasks. For text-to-image generation, we employ classifier-free guidance (CFG) [20] with a guidance weight set to 5. More details on the training data and settings are provided in App. A.

4.2 Quantitative Results

Text-to-Image Generation. Table 1 presents a comprehensive comparison between ULM-R1 and state-of-the-art models across three visual generation benchmarks. Among unified models, our model ranks second on both GenEval and WISE benchmarks. Notably, it achieves balanced performance across diverse task categories within GenEval, with the best score of 0.71 in object counting. When compared with specialized generation-only models, ULM-R1 surpasses the top performer SD3-Medium [15] by a slight margin (0.77 *vs.* 0.74 on GenEval). Moreover, ULM-R1 shows consistent improvements over its base model across all benchmarks. These results collectively demonstrate the effectiveness and advantage of our CoRL in enhancing visual generation quality.

Multimodal Understanding. Results are shown in Table 2. For mixed QA format evaluation, we continue to apply the Gaussian-distribution-based merging strategy [66] to combine the two task-specific policy models and obtain a final model capable of following both types of instructions. Overall, ULM-R1 markedly outperforms existing unified models across most benchmarks, and substantially narrows the performance gap with leading understanding-only MLLMs of comparable model scale. More specifically, our model achieves state-of-the-art performance among unified models

Table 3: **Comparison between different RL paradigms for ULMs.** The cold SFT data is consist of `x2x_rft_22k`, `mcot_r1_mcq` (22K), and `mcot_r1_vqa` (22K). #7: CoRL.

#	Ablated Setting	Stage	GenEval	DPG	MMMU	Math^{We}	MMVet	Logic^{VT}
0	Baseline	-	73.0	82.6	36.3	5.9	39.8	23.9
1	+ Cold-SFT	S1	72.8 (-0.3)	82.5 (-0.1)	41.0 (+4.7)	18.0 (+12.1)	42.0 (+2.2)	27.9 (+4.0)
2	+ Unified-RL	S1	75.9 (+2.9)	83.3 (+0.7)	40.3 (+4.0)	14.0 (+8.1)	42.5 (+2.7)	30.2 (+6.3)
3	+ Refined-RL (T2I)	S2	75.1 (+2.1)	83.0 (+0.4)	/	/	/	/
4	+ Refined-RL (MM2T-MC)	S2	/	/	39.6 (+3.3)	15.8 (+9.9)	/	/
5	+ Refined-RL (MM2T-OE)	S2	/	/	/	/	42.2 (+2.4)	29.5 (+5.6)
6	+ Refined-RL w/ Cold-SFT	S1&S2	74.5 (+1.5)	82.8 (+0.2)	41.8 (+5.5)	22.5 (+16.6)	43.7 (+3.9)	35.9 (+12.0)
7	+ Refined-RL w/ Unified-RL	S1&S2	77.3 (+4.3)	83.9 (+1.3)	42.3 (+6.0)	21.1 (+15.2)	43.9 (+4.1)	34.5 (+10.6)

Figure 3: **Qualitative comparison of text-to-image generation** between Janus-Pro and ULM-R1. The red box marks an exemplary failure case.

on MMStar (47.6), WeMath (21.1), LogicVista (34.5), and on several mixed-format math benchmarks, including MathVerse (25.4) and MathVision (22.0). Particularly, ULM-R1 demonstrates considerable improvements over its base model in mathematical and logical reasoning tasks, achieving gains of **15.2** on WeMath and **10.6** on LogicVista. These results not only demonstrate the effectiveness of CoRL in enhancing ULMs' understanding capabilities, but also establish that reinforcement learning provides a data-efficient pathway for achieving both robust generalization and sophisticated reasoning capabilities, without the need for large-scale supervised data.

4.3 Qualitative Results

In this section, we first present a qualitative comparison between ULM-R1 and Janus-Pro for visual generation, as illustrated in Figure 3. The results clearly show that ULM-R1 achieves superior text-to-image alignment and object grounding across diverse prompts, with especially notable improvements in spatial arrangement of objects and compositional consistency. Next, as shown in Figure 4, we visualize several representative examples of multimodal understanding. Compared to Janus-Pro, ULM-R1 exhibits significantly enhanced understanding capabilities, particularly in mathematical reasoning. These comprehensive qualitative results demonstrate the effectiveness of CoRL in simultaneously

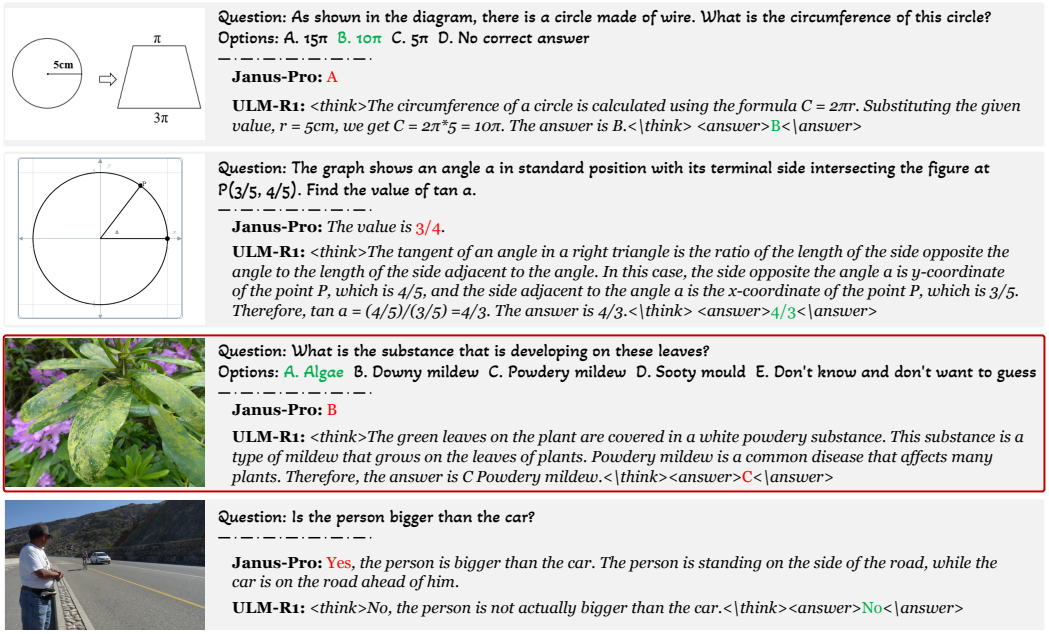

Figure 4: **Qualitative comparison of multimodal understanding** between Janus-Pro and ULM-R1. The red box marks an exemplary failure case.

improving visual generation and multimodal understanding in ULMs. Furthermore, Figures 3 and 4 respectively showcase exemplary failure cases of ULM-R1 in visual generation and understanding tasks, providing an intuitive grasp of its limitations. For instance, in the understanding example, it misinterprets commonsense and professional knowledge, leading to an incorrect answer.

4.4 Ablation Study and Discussion

In this section, we primarily evaluate the effectiveness of our RL training strategy and the proposed reward functions for text-to-image generation. In addition, we discuss the impact of the hyperparameter λ and the scalability of CoRL.

Comparison Between Various RL Paradigms. As presented in Table 3, we conduct a comprehensive ablation study to evaluate the effects of different RL paradigms for ULMs. The results reveal two key findings: ▶ #2 *vs.* #1: Unified-RL effectively enhances both the generation and understanding capabilities of ULMs, whereas Cold-SFT has minimal impact on visual generation. ▶ #7 *vs.* #6: Compared to the de facto paradigm, our CoRL consistently outperforms it on visual generation benchmarks while achieving comparable results on multimodal understanding benchmarks. These findings indicate that *unified RL provides a robust foundation for task-specific refinement, even without reliance on supervised data*. Additionally, CoRL consistently outperforms both its baseline and task-specific RL variants (#3-#5), achieving improvements of 2.1 points on GenEval (*vs.* generation-only RL, #3) and 5.3 points on WeMath (*vs.* understanding-only RL, #4). These results demonstrate the efficacy of CoRL as our final RL paradigm.

Effect of Rewards in Text-to-Image Generation. To evaluate the effectiveness of our proposed rewards for text-to-image generation, we conduct ablation experiments as detailed in Table 4. The results demonstrate that incorporating either reward individually improves performance over the baseline: $\mathcal{R}_{\text{cycle}}$ yields an increase of 2.1 in average score, while $\mathcal{R}_{\text{TIM}}$ results in an increase of 0.8. Notably, combining both rewards leads to the best overall performance, achieving an average score of 80.6. These findings suggest a modest but complementary effect between $\mathcal{R}_{\text{cycle}}$ and $\mathcal{R}_{\text{TIM}}$, enhancing their joint benefit in enhancing visual generation quality. In addition, we further compare the CLIP score ($\mathcal{R}_{\text{CLIP}}$) and $\mathcal{R}_{\text{TIM}}$ under our final RL training paradigm. As shown in the table, $\mathcal{R}_{\text{TIM}}$ achieves better overall performance, especially on the DPG benchmark with dense, long-horizon prompts for image generation, highlighting its superior ability to capture fine-grained semantic alignment compared to the CLIP score.

Table 4: Effect of visual generation rewards.

Rewards	GenEval	DPG	Avg. ↑
Baseline	73.0	82.6	77.8
$\mathcal{R}_{\text{CLIP}}$	74.2	82.4	78.3 (+0.5)
$\mathcal{R}_{\text{TIM}}$	74.1	83.0	78.6 (+0.8)
$\mathcal{R}_{\text{cycle}}$	76.2	83.5	79.9 (+2.1)
$\mathcal{R}_{\text{cycle}}$ + $\mathcal{R}_{\text{CLIP}}$	77.0	83.4	80.2 (+2.4)
$\mathcal{R}_{\text{cycle}}$ + $\mathcal{R}_{\text{TIM}}$	77.3	83.9	80.6 (+2.8)

Table 5: Comparison among visual consistency measures used in $\mathcal{R}_{\text{cycle}}$.

Measures	GenEval	DPG
PSNR	76.0	82.4
MSE	77.1	83.2
SSIM	77.5	83.6
LPIPS	77.3	83.9

Table 6: Impact of λ.

λ	GenEval (Gen.)	MMMU (Und.)
0.5	77.1	41.0
0.7	77.3	42.3
0.8	77.3	42.3
0.9	77.1	43.5
1.0	76.9	43.0

Impact of Visual Consistency Measures in $\mathcal{R}_{\text{cycle}}$. Table 5 provides a more detailed analysis of how different visual consistency measures (PSNR, MSE, SSIM, and LPIPS) used in $\mathcal{R}_{\text{cycle}}$ affect the quality of visual generation. PSNR and MSE are pixel-level metrics that quantify low-level differences between the generated images, while SSIM and LPIPS assess higher-level perceptual and structural similarities. As shown in the table, SSIM and LPIPS perform better than the other two metrics, with LPIPS achieving the best performance (83.9) on the DPG benchmark. This can be attributed to the fact that LPIPS measures image similarity in a feature space, making it more robust to minor, semantically irrelevant variations and thus better suited to reward high-level consistency.

Impact of hyperparameter λ. The factor λ in Eq. (5) balances the reward scales between the two tasks during unified RL. As shown in Table 6, we conduct experiments using different values of λ to assess its impact on both generation and understanding performance. The results show that moderate values of λ (~ 0.8) achieve a balanced trade-off between generation and understanding. Larger values slightly degrade generation performance, indicating that overemphasizing understanding rewards may hinder cross-task optimization.

Scalability of CoRL. To validate the effectiveness of CoRL on other ULMs, as illustrated in Table 7, we conduct additional experiments using Janus-1.3B [85] and Janus-Pro-7B [8] as the baseline. The results show consistent improvements across both generation and understanding benchmarks, confirming the scalability of CoRL. Notably, Janus-Pro-7B with LoRA tuning achieves smaller gains on the mathematical reasoning benchmark (WeMath) than Janus-1.3B, suggesting that while CoRL scales well across model size, its enhancement of complex reasoning does not scale linearly.

Table 7: **Effectiveness of CoRL on other ULMs.** For Janus-Pro-7B, we adopt LoRA tuning [21] to enable efficient training and mitigate memory pressure during unified RL.

Methods	GenEval	WISE	DPG	MMMU	MMStar	Math^{We}	MMVet	POPE	Logic^{VT}
Janus-1.3B	0.61	0.23	79.68	30.5	37.6	3.4†	34.3	87.0	23.9
+ CoRL (Full Fine-Tuning)	0.64	0.26	80.92	34.6	41.9	16.4	36.9	88.1	27.0
Janus-Pro-7B	0.80	0.35	84.19	41.0	46.5	9.7	50.0	87.4	28.0
+ CoRL (LoRA Tuning)	0.82	0.41	84.97	44.6	49.5	16.0	52.6	88.0	32.4

5 Limitation

Despite the substantial improvements achieved, several limitations remain that warrant further investigation. First, a notable performance gap still exists between generation and understanding tasks of ULMs. Second, our rewards for multimodal understanding are relatively simple and primary. These limitations highlight the need for more sophisticated RL designs that can further enhance understanding capabilities and narrow the performance gap. We hope our work provides valuable insights for future RL research in ULMs.

6 Conclusion

In this work, we investigate how to jointly enhance the understanding and generation capabilities of ULMs, and propose a co-reinforcement learning framework (CoRL). Within the proposed CoRL, the policy model follows a Foundation-then-Specialization paradigm that involves a two-stage RL procedure: a unified RL stage for joint optimization and a refined RL stage for task-specific enhancement, yielding ULM-R1. Extensive evaluations across diverse understanding and generation benchmarks demonstrate the effectiveness of CoRL and the advantage of ULM-R1.

Acknowledgements. This work was supported in part by NSFC (62406189, 62322113, 62376156), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and the Fundamental Research Funds for the Central Universities.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398, 2016. 4

[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. 7

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions, 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf. 6

[4] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv:2505.09568*, 2025. 2

[5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv:2310.00426*, 2023. 6

[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? *arXiv:2403.20330*, 2024. 6

[7] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M^3CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv:2405.16473*, 2024. 24

[8] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv:2501.17811*, 2025. 1, 2, 3, 6, 7, 10

[9] Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *arXiv:2503.07065*, 2025. 1

[10] Linger Deng, Yuliang Liu, Bohan Li, Dongliang Luo, Liang Wu, Chengquan Zhang, Pengyuan Lyu, Ziyang Zhang, Gang Zhang, Errui Ding, et al. R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multimodal models. *arXiv:2410.17885*, 2024. 24

[11] Hongyuan Dong, Zijian Kang, Weijie Yin, Xiao Liang, Chao Feng, and Jiao Ran. Scalable vision language model training via high quality data curation. *arXiv:2501.05952*, 2025. 7

[12] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. In *ICLR*, 2024. 2

[13] 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. *arXiv:2407.11691*, 2024. 7

[14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 2

[15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 6, 7

[16] Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, et al. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv:2503.13436*, 2025. 2

[17] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv:2404.14396*, 2024. 2

[18] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, pages 52132–52152, 2023. 6

[19] 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, 3, 5

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022. 7

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 10

[22] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv:2403.05135*, 2024. 6

[23] Runhui Huang, Chunwei Wang, Junwei Yang, Guansong Lu, Yunlong Yuan, Jianhua Han, Lu Hou, Wei Zhang, Lanqing Hong, Hengshuang Zhao, et al. Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. *arXiv:2504.01934*, 2025. 2, 6, 7

[24] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv:2503.06749*, 2025. 1, 3

[25] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv:2505.00703*, 2025. 1

[26] Jingjing Jiang, Chao Ma, Xurui Song, Hanwang Zhang, and Jun Luo. Corvid: Improving multimodal large language models towards chain-of-thought reasoning. *arXiv:2507.07424*, 2025. 24

[27] Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. *arXiv:2504.04423*, 2025. 7

[28] Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv:2312.12241*, 2023. 24

[29] Siqi Kou, Jiachun Jin, Chang Liu, Ye Ma, Jian Jia, Quan Chen, Peng Jiang, and Zhijie Deng. Orthus: Autoregressive interleaved image-text generation with modality-specific heads. *arXiv:2412.00127*, 2024. 2, 6, 7

[30] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. Llm post-training: A deep dive into reasoning large language models. *arXiv:2502.21321*, 2025. 1

[31] Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, et al. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. *arXiv:2412.09604*, 2024. 2, 7

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742, 2023. 4

[33] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv:2403.00231*, 2024. 24

[34] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment. *arXiv:2410.09421*, 2024. 3

[35] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv:2504.06958*, 2025. 3

[36] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv:2305.10355*, 2023. 7

[37] Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. *arXiv:2501.00289*, 2024. 6

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 3, 24

[39] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. In *ICLR*, 2025. 1, 2, 6

[40] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv:2503.06520*, 2025. 1, 3

[41] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv:2503.01785*, 2025. 1, 3

[42] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, pages 2507–2521, 2022. 24

[43] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv:2310.02255*, 2023. 7

[44] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *ICLR*, 2023. 24

[45] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*, 2024. 7

[46] Xingtai Lv, Yuxin Zuo, Youbang Sun, Hongyi Liu, Yuntian Wei, Zhekai Chen, Lixuan He, Xuekai Zhu, Kaiyan Zhang, Bingning Wang, et al. Towards a unified view of large language model post-training. *arXiv:2509.04419*, 2025. 1

[47] Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv:2502.20321*, 2025. 2

[48] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Liang Zhao, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv:2411.07975*, 2024. 2

[49] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv:2504.05299*, 2025. 7

[50] Yuwei Niu, Munan Ning, Mengren Zheng, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv:2503.07265*, 2025. 6

[51] OpenAI. Openai o1 system card. *arXiv:2412.16720*, 2024. 1

[52] OpenAI. Openai o3 and o4-mini system card, 2025. URL https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf. 1

[53] 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. 1, 7

[54] Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. Strengthening multimodal large language model with bootstrapped preference optimization. In *ECCV*, pages 382–398, 2024. 3

[55] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv:2307.01952*, 2023. 6

[56] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv:2407.01284*, 2024. 7

[57] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv:2412.03069*, 2024. 2, 6

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3, 5

[59] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, pages 53728–53741, 2023. 1, 3

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 6

[61] Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. Mitigating object hallucination via data augmented contrastive tuning. *arXiv:2405.18654*, 2024. 3

[62] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017. 3

[63] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *ECCV*, pages 146–162, 2022. 24

[64] 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, 3

[65] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation. *arXiv:2412.15188*, 2024. 2

[66] Chongjie Si, Jingjing Jiang, and Wei Shen. Unveiling the mystery of weight in large foundation models: Gaussian distribution never fades. *arXiv:2501.10661*, 2025. 4, 7

[67] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv:2406.06525*, 2024. 2, 6

[68] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv:2309.14525*, 2023. 3

[69] 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. 1, 3

[70] Yunlong Tang, Jing Bi, Pinxin Liu, Zhenyu Pan, Zhangyun Tan, Qianxiang Shen, Jiani Liu, Hang Hua, Junjia Guo, Yunzhong Xiao, et al. Video-lmm post-training: A deep dive into video reasoning with large multimodal models. *arXiv:2510.05034*, 2025. 1

[71] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv:2405.09818*, 2024. 2

[72] 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

[73] Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, et al. A survey on post-training of large language models. *arXiv:2503.06072*, 2025. 1

[74] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv:2412.14164*, 2024. 2

[75] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020. 7

[76] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. *arXiv:2412.06673*, 2024. 2

[77] Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. *arXiv:2406.11839*, 2024. 1, 3

[78] Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv:1611.05763*, 2016. 3

[79] Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. *arXiv:2504.11455*, 2025. 1, 3, 6

[80] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv:2402.14804*, 2024. 7

[81] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv:2411.10442*, 2024. 3

[82] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv:2409.18869*, 2024. 2, 3, 6

[83] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv:2503.05236*, 2025. 1, 3

[84] Marco A Wiering and Martijn Van Otterlo. Reinforcement learning. *Adaptation, learning, and optimization*, 12(3):729, 2012. 3

[85] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv:2410.13848*, 2024. 2, 6, 7, 10

[86] Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable multi-modal generators. *arXiv:2412.04332*, 2024. 2

[87] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv:2503.21979*, 2025. 6, 7

[88] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. In *ICLR*, 2025. 1, 2, 7

[89] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv:2407.04973*, 2024. 7

[90] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *ICLR*, 2025. 1, 2, 6, 7

[91] Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. *arXiv:2411.17762*, 2024. 2

[92] Ling Yang, Xinchen Zhang, Ye Tian, Chenming Shang, Minghao Xu, Wentao Zhang, and Bin Cui. Hermesflow: Seamlessly closing the gap in multimodal understanding and generation. *arXiv:2502.12148*, 2025. 2, 3, 6, 7

[93] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv:2110.04627*, 2021. 2

[94] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv:2503.14476*, 2025. 5

[95] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *CVPR*, pages 13807–13816, 2024. 1, 3

[96] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv:2405.17220*, 2024. 3

[97] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv:2308.02490*, 2023. 7

[98] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, pages 9556–9567, 2024. 6

[99] Simon Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Peter Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, et al. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. In *NeurIPS*, pages 110935–110971, 2024. 3

[100] Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *arXiv:2503.18013*, 2025. 1, 3

[101] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv:2503.12937*, 2025. 1

[102] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv:2403.14624*, 2024. 7

[103] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv:2407.08739*, 2024. 24

[104] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 4

[105] Zhiyuan Zhao, Linke Ouyang, Bin Wang, Siyuan Huang, Pan Zhang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Mllm-dataengine: An iterative refinement approach for mllm. *arXiv:2308.13566*, 2023. 24

[106] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv:2311.16839*, 2023. 3

[107] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv:2408.11039*, 2024. 2

[108] Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. *arXiv:2504.21277*, 2025. 1, 3

[109] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv:2402.11411*, 2024. 3

[110] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv:2504.10479*, 2025. 7

[111] Jialv Zou, Bencheng Liao, Qian Zhang, Wenyu Liu, and Xinggang Wang. Omnimamba: Efficient and unified multimodal understanding and generation via state space models. *arXiv:2503.08686*, 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: The main claims made in the abstract and introduction accurately reflect our paper's contributions and scope.

   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: We discuss the limitations in the last section of the paper.

   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 results.

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: We provide all implementation details.

   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: [Yes]

Justification: We provide all links to our data in the footnote of the paper and will soon release our code.

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: We provide all the experimental details.

   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: [No]

   Justification: We use official evaluation protocols provided by the corresponding benchmarks.

   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: We provide all the implementation details.

   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: We have checked our work.

   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: We discuss this in the conclusion and introduction sections.

    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 paper poses no such risks.

    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: All are properly referred to.

    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: The paper does not release new assets.

    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: No such experiments.

    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: No such experiments.

    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: LLM is used for paper polishing.

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.

A Appendix

A.1 Training Data

Training Data for Unified Reinforcement Learning. To support synergistic multimodal modeling during unified RL, we curate a dataset (*i.e.*, `x2x_rft_22k`) that simultaneously involves text-to-image generation and multimodal understanding tasks. As illustrated in Figure 5, each sample includes *a real image*, *a prompt* for generation, and *a problem* for understanding. The real images are sourced from the COCO 2017 train split [38], while the problems and their corresponding solutions are adapted from A-OKVQA [63] and GPT-VQA [105]. In addition, prompts are selected from the original COCO captions based on their entity coverage with the problem solutions.

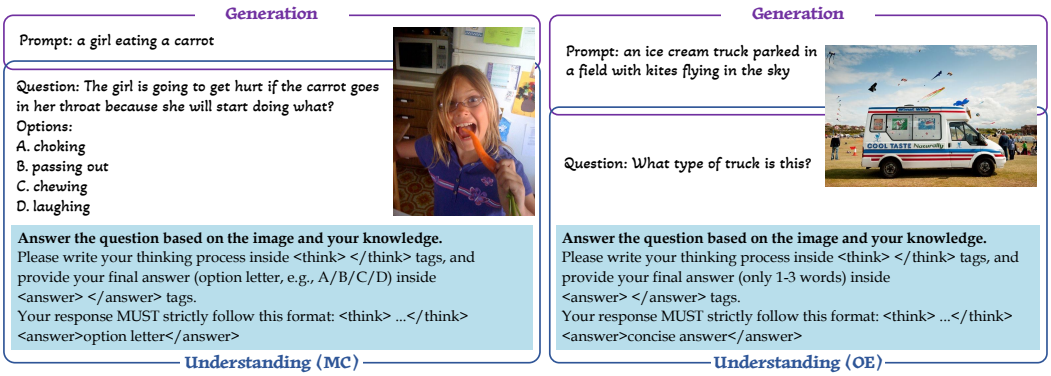

Figure 5: **Illustration of training examples used in unified reinforcement learning.**

Training Data for Refined Reinforcement Learning. In this stage, we collect three specialized datasets for task-specific RL. For text-to-image generation, we continue constructing a dataset (*i.e.*, `x2x_rft_16k`) with prompts derived from COCO captions. Moreover, we curate `mcot_r1_mcq` and `mcot_r1_vqa` for multiple-choice and open-ended multimodal understanding, respectively. These two datasets are curated on top of MCoT-Instruct [26], which encompasses a diverse range of multimodal tasks, including mathematical reasoning, science-problem solving, and visual commonsense reasoning, across multiple source datasets. Specifically, the source datasets of `mcot_r1_mcq` comprise A-OKVQA [63], M^3CoT [7], SQA-IMG (train) [42], ArxivQA [33], TabMWP (MC) [44], and MAVIS-Instruct (MC) [103], while the source datasets of `mcot_r1_vqa` include GeomVerse [28], R-CoT [10], TabMWP (OE) [44], and MAVIS-Instruct (OE) [103].

A.2 Supplementary Experimental Setups

Table 8 provides detailed hyperparameter settings for ULM-R1's RL training.

Table 8: **Training hyperparameter setting.**

Configuration	Unified RL	Refined RL (T2I)	Refined RL (MM2T-MC)	Refined RL (MM2T-OE)
Number of sampled outputs (G)	8	16	16	16
Regularization coefficient of $\mathbb{D}_{\text{KL}}$ (β)	0	0.02	0.02	0.02
Max prompt length	1024	256	1024	1024
Max completion length	512	/	512	512
Batch size	16	16	32	32
Peak learning rate	4e-6	1e-6	1e-6	1e-6
Epoch	1	1	1	1