NEURAL INFORMATION PROCESSING SYSTEMS

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

■ DeepSeek-R1

◉ The RL with verifiable rewards and GRPO is a promising post-training paradigm, enabling pretrained LFMs effectively acquire advanced capabilities and generalization without dependence on large-scale, high-quality supervised data

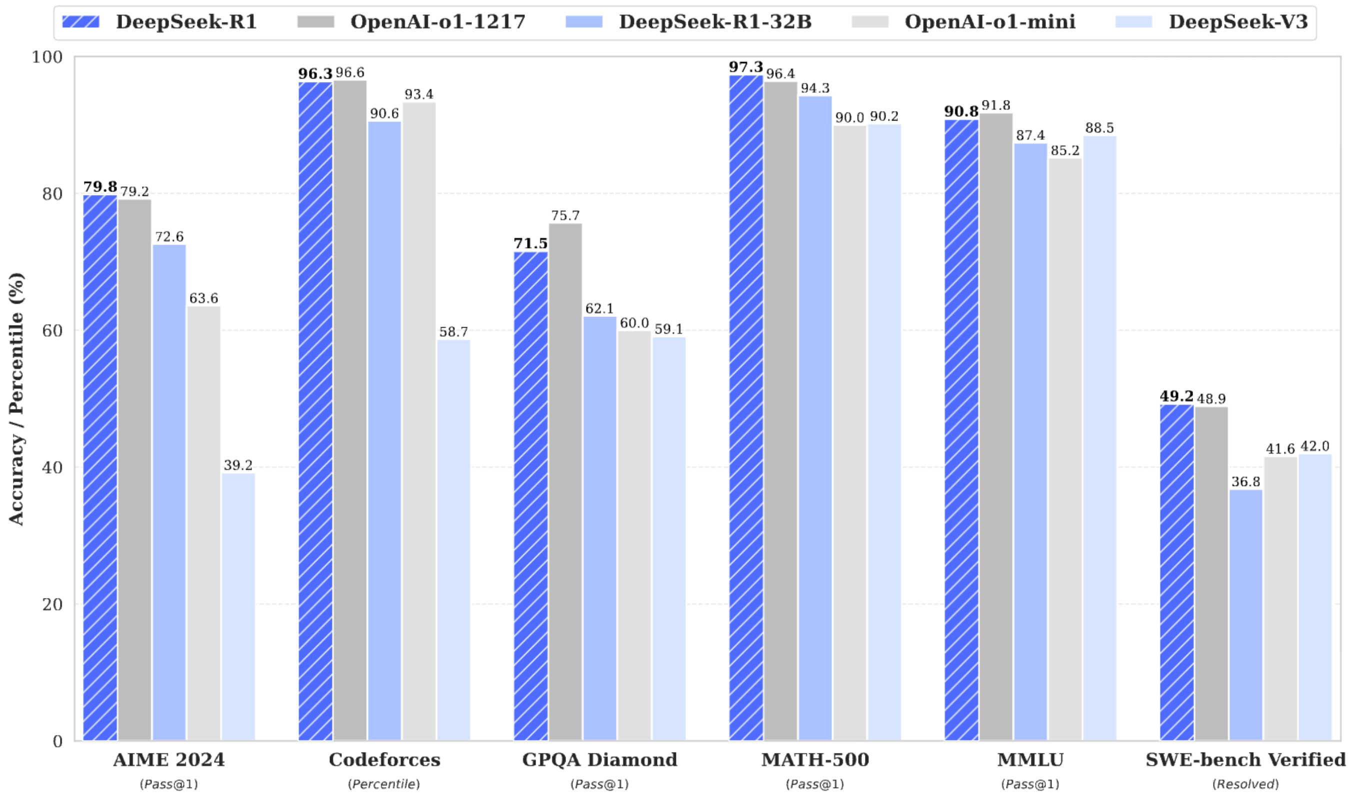


Figure 1 | Benchmark performance of DeepSeek-R1.

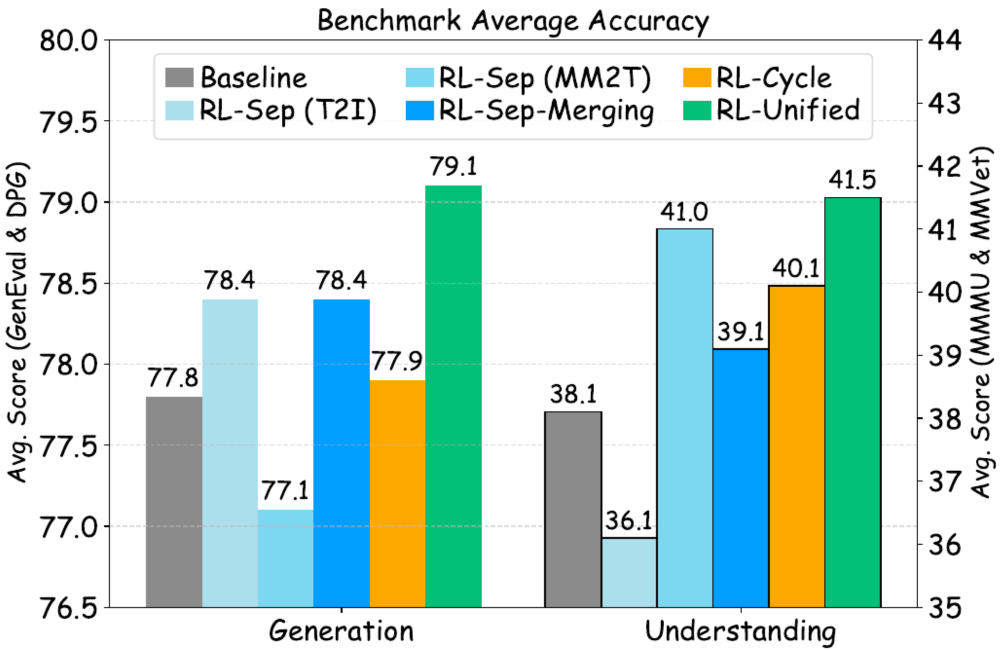
Daya Guo, Dejian Yang, et al. Deepseek-R1: Incentivizing reasoning capability in llms via reinforcement learning. DeepSeek, 2025.

Motivation

■ Pilot Exploration

◉ **Comparing 4 distinct RL paradigms using Janus-Pro-1B as the baseline ULMs across both generation and understanding tasks**

- ✓ **RL-Sep:** understanding and generation tasks are independently optimized under their respective rewards
- ✓ **RL-Merging:** separate RL followed by weight merging strategy to incorporate both abilities
- ✓ **RL-Cycle**: using a scheduled alternation between the two tasks throughout the training process
- ✓ **RL-Unified:** both tasks are jointly optimized within a unified framework to promote co-evolution



Findings:

① Direct single-task RL fails to achieve the expected improvements for ULMs, particularly in the visual generation task, and may even impair the other task's performance

② Compared with alternative strategies, unified RL demonstrates average performance advantages over alternative paradigms

⇩

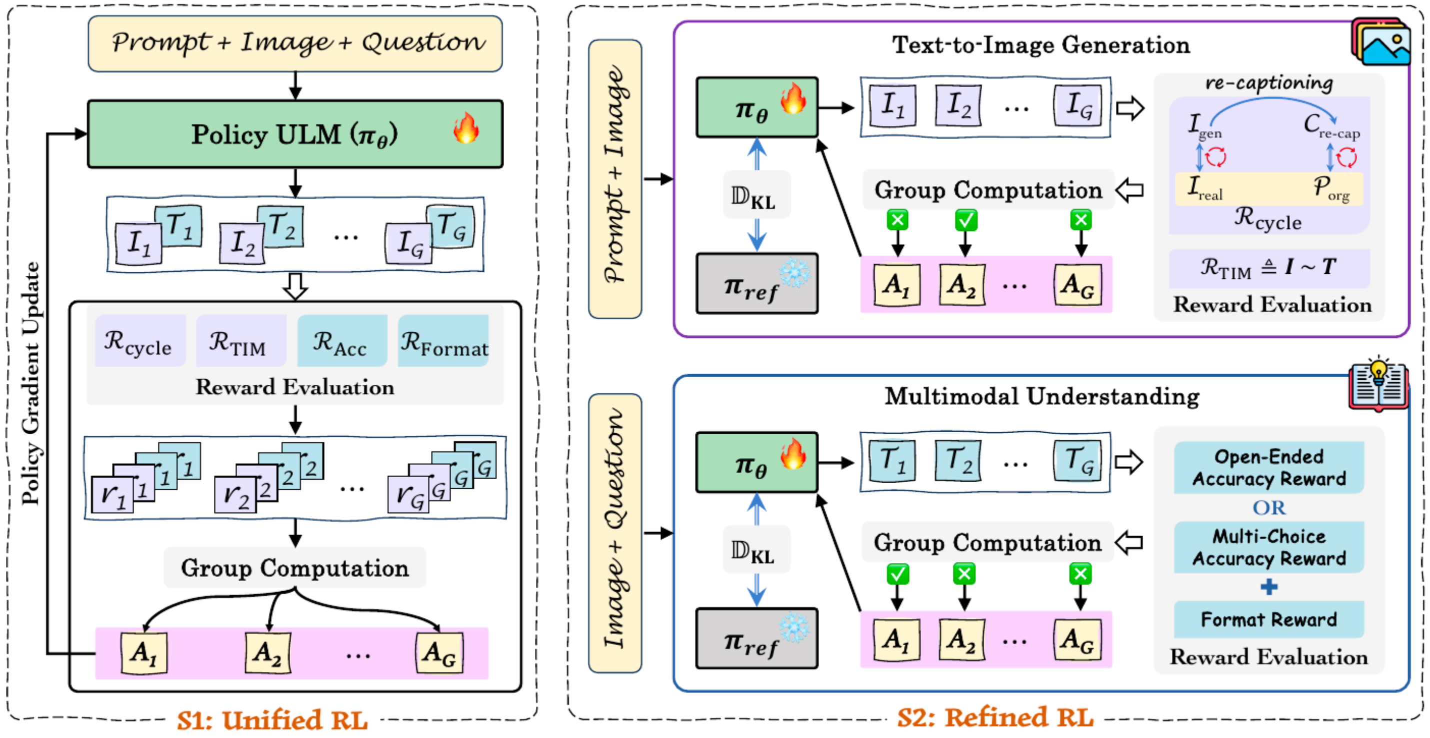
The synergistic co-evolution of dual capabilities under a shared policy optimization paradigm

Methodology

■ Co-Reinforcement Learning (CoRL)

A two-stage (*unified-then-refined*) RL paradigm

- ➢ Unified RL (Stage 1): to jointly optimize the dual capabilities of ULMs and build a powerful generalist foundation
- ➢ Refined RL (Stage 2): to further improve target tasks built upon the established strong foundation



Verifiable Rewards

① **Bidirectional Cycle Consistency Reward**

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

② **Text-Image Matching Reward**

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

③ **Accuracy Reward**

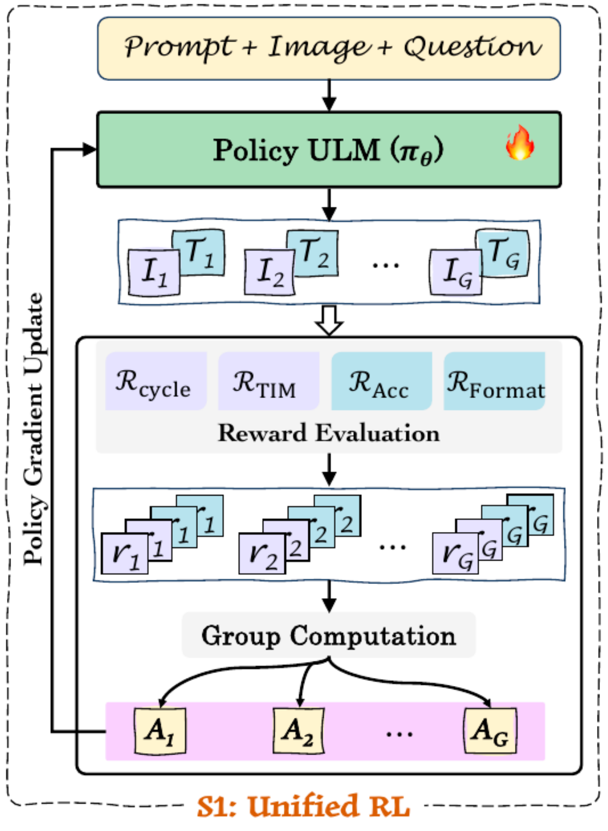
$\mathcal{R}_{\text{MCQ-Acc}}$ $\mathcal{R}_{\text{OE-Acc}}$

④ **Format Reward**

$\mathcal{R}_{\text{Format}}$

Methodology

■ Unified RL (Stage 1)



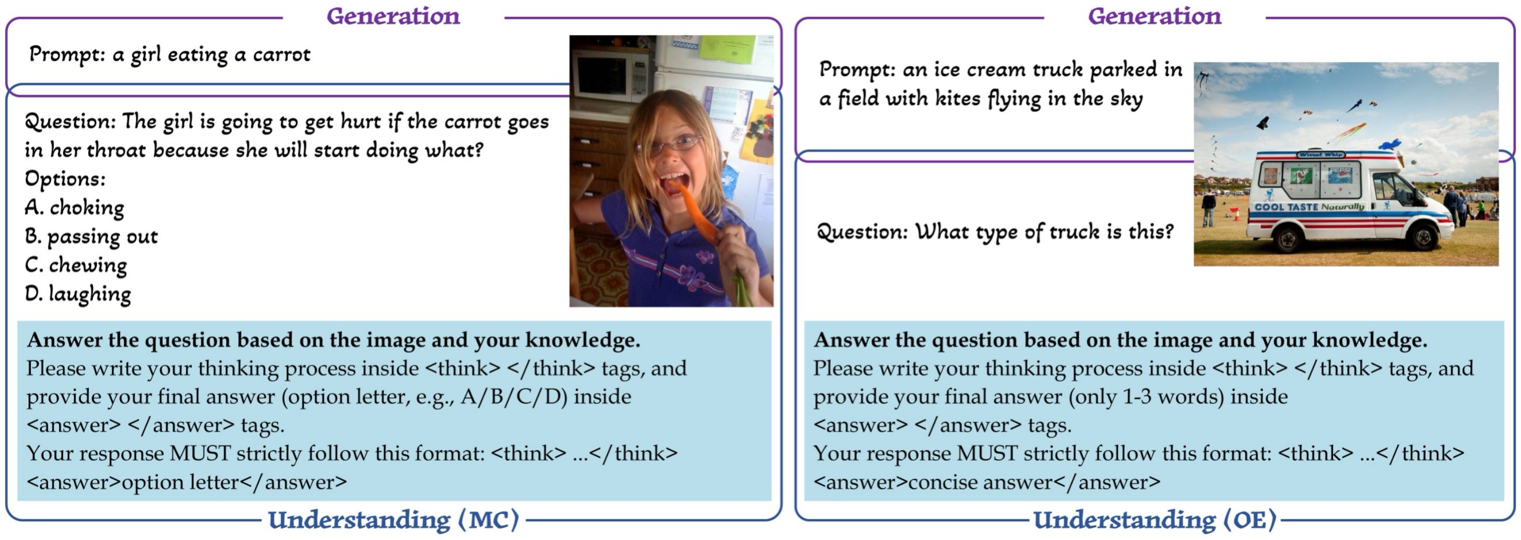
Reward Function

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

Training Objective

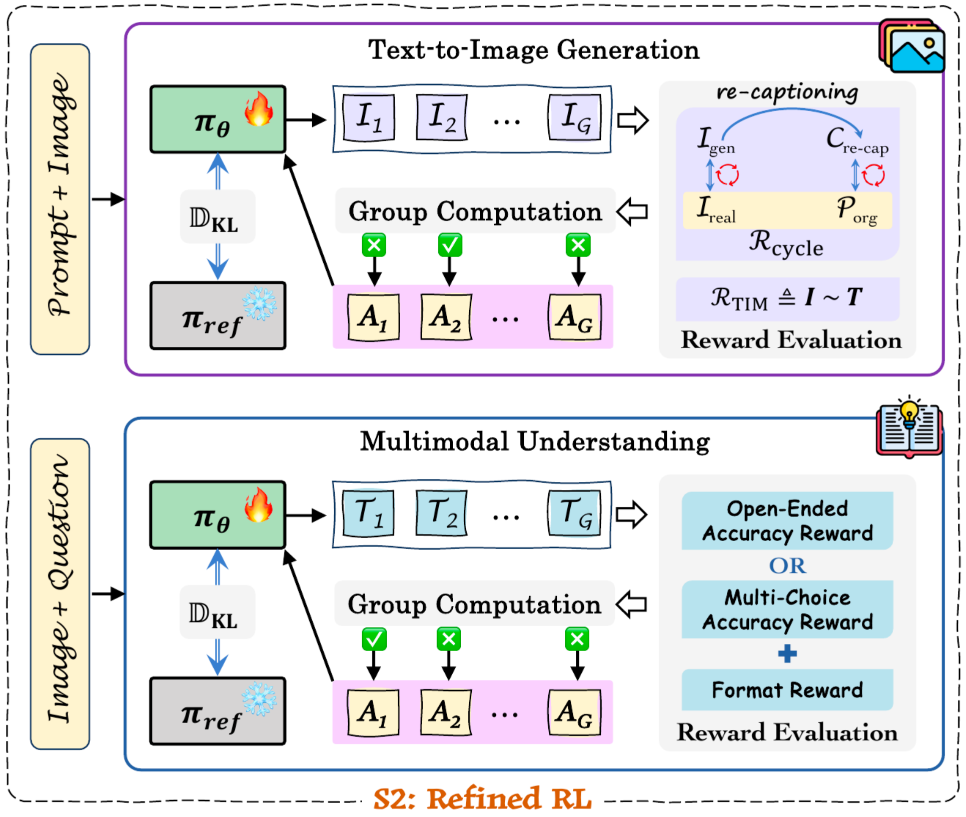
$$\mathcal{L}_{\text{S1}} = \mathbb{E}_{\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}} \frac{1}{G} \sum_{i=1}^{G} \frac{\pi_{\theta}(o_i)}{\pi_{\theta_{\text{old}}}(o_i)} A_i \, , \text{ where } o_i = (\mathcal{I}_i, \mathcal{T}_i)$$

Training Data Format: Triplet <Image, Prompt, Question>



Methodology

■ Refined RL (Stage 2)



Reward Function for Text-to-Image Generation

$$\mathcal{R}_{\text{T2I-S2}} = \mathcal{R}_{\text{cycle}} + \mathcal{R}_{\text{TIM}}$$

Reward Function for Multimodal Understanding

$$\mathcal{R}_{\text{MCQ-S2}} = \mathcal{R}_{\text{MCQ-Acc}} + \mathcal{R}_{\text{Format}} \qquad \mathcal{R}_{\text{OE-S2}} = \mathcal{R}_{\text{OE-Acc}} + \mathcal{R}_{\text{Format}}$$

Training Objective

$$\mathcal{L}_{S2} = \mathbb{E}_{\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}} \frac{1}{G} \sum_{i=1}^{G} \left[\frac{\pi_{\theta}(o_i)}{\pi_{\theta_{\text{old}}}(o_i)} A_i - \beta \, \mathbb{D}_{\text{KL}} \left(\pi_{\theta} \,\|\, \pi_{\text{ref}} \right) \right]$$

Experiment Results

■ Text-to-Image Generation

Model	Scale	Res.	Type	GenEval ↑ Two Obj.	Counting	Position	Color Attri.	Overall	WISE ↑ Overall	DPG ↑ 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 [58]	0.9B	512^2	Diff	0.38	0.35	0.04	0.06	0.43	0.32	63.18
SDv2.1 [58]	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 [53]	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 [65]	0.8B	256^2	F-AR	0.34	0.21	0.07	0.04	0.32	-	65.16
SimpleAR [76] ♣	1.5B	1024^2	F-AR	0.90	-	0.28	0.45	0.63	-	81.97
▼ *Unified Understanding and Generation*										
TokenFlow [55]	8B	256^2	F-AR	0.60	0.41	0.16	0.24	0.55	-	73.38
Emu3 [79]	8B	512^2	F-AR	-	-	-	-	0.66	0.39	80.60
Emu3-DPO [79] ♣	8B	512^2	F-AR	-	-	-	-	0.64	-	81.60
LWM [38]	7B	512^2	F-AR	0.41	0.46	0.09	0.15	0.47	-	-
Orthus [28]	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+ [22]	3B	384^2	AR-Diff	0.88	0.62	0.42	0.53	0.72	-	-
D-DiT [36]	2B	512^2	Diff	0.80	0.54	0.32	0.50	0.65	-	-
Harmon [84]	1.5B	512^2	F-AR	0.86	0.66	0.74	0.48	0.76	**0.41**	-
show-o [87]	1.3B	512^2	AR-Diff	0.80	0.66	0.31	0.50	0.68	0.35	67.48
HermesFlow [89] ♣	1.3B	512^2	AR-Diff	0.84	0.66	0.32	0.52	0.69	-	70.22
Janus [82]	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

Experiment Results

- **Text-to-Image Generation:** ULM-R1 demonstrates superior text-to-image alignment and object grounding across diverse prompts, with especially notable improvements in spatial arrangement of objects and compositional consistency

failure case

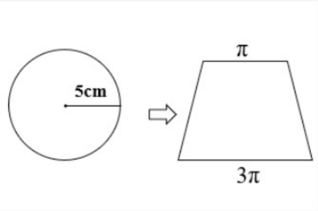
Experiment Results

▪ Multimodal Understanding

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 [47]	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 [44]	Qwen2.5-1.5B	45.6	56.7	9.9	58.3	87.8	34.7	**64.1**	29.4	17.7
InternVL3 [107]	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 [51]	Qwen2.5-3B	-	58.0	-	-	-	-	63.2	**41.6**	**26.4**
▼ *Unified Understanding and Generation*										
ILLUME+ [22]	Qwen2.5-3B	**44.3**	-	-	40.3	87.6	-	-	-	-
Harmon [84]	Qwen2.5-1.5B	38.9	-	-	-	87.6	-	-	-	-
VILA-U [85]	LLaMA-2-7B	-	-	-	33.5	85.8	-	-	-	-
Orthus [28]	Chameleon-7B	28.2	-	-	-	79.6	-	-	-	-
UniToken [26]	Chameleon-7B	32.8	46.1	-	-	-	-	38.5	-	
SGen-VL [30]	InternLM2-1.8B	34.2	-	-	34.5	85.3	-	**42.7**	-	
Show-o [87]	Phi-1.3B	26.7	-	-	-	80.0	-	-	-	-
HermesFlow [89]	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 [82]	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**

Experiment Results

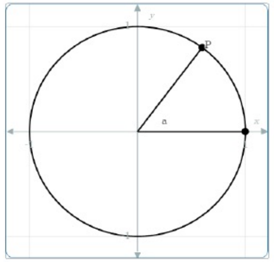
Multimodal Understanding: ULM-R1 exhibits significantly enhanced understanding capabilities, particularly in mathematical reasoning



Question: As shown in the diagram, there is a circle made of wire. What is the circumference of this circle?
Options: A. 15π B. 10π C. 5π D. No correct answer

Janus-Pro: A

ULM-R1: *<think>The circumference of a circle is calculated using the formula $C = 2\pi r$. Substituting the given value, r = 5cm, we get $C = 2\pi * 5 = 10\pi$. The answer is B.<\think> <answer>B<\answer>*



Question: The graph shows an angle a in standard position with its terminal side intersecting the figure at P(3/5, 4/5). Find the value of tan a.

Janus-Pro: *The value is 3/4.*

ULM-R1: *<think>The tangent of an angle in a right triangle is the ratio of the length of the side opposite the angle to the length of the side adjacent to the angle. In this case, the side opposite the angle a is y-coordinate of the point P, which is 4/5, and the side adjacent to the angle a is the x-coordinate of the point P, which is 3/5. Therefore, tan a = (4/5)/(3/5) =4/3. The answer is 4/3.<\think> <answer>4/3<\answer>*

Question: What is the substance that is developing on these leaves?
Options: A. Algae B. Downy mildew C. Powdery mildew D. Sooty mould E. Don't know and don't want to guess

Janus-Pro: B

ULM-R1: *<think>The green leaves on the plant are covered in a white powdery substance. This substance is a type of mildew that grows on the leaves of plants. Powdery mildew is a common disease that affects many plants. Therefore, the answer is C Powdery mildew.<\think><answer>C<\answer>*

failure case

Question: Is the person bigger than the car?

Janus-Pro: Yes, *the person is bigger than the car. The person is standing on the side of the road, while the car is on the road ahead of him.*

ULM-R1: *<think>No, the person is not actually bigger than the car.<\think><answer>No<\answer>*

Conclusion

- In this work, we proposed a **co-reinforcement learning framework (CoRL)** to jointly enhance the understanding and generation capabilities of ULMs.

- CoRL implements a two-stage (unified-then-refined) RL paradigm to jointly improve the understanding and generation capabilities of ULMs.

- Despite the substantial improvements achieved, several limitations remain, such as **the capability gap between understanding and generation** and **the primitive rewards for understanding**, which warrant further investigation.