CVPR JUNE 3-7, 2026 DENVER COLORADO

CodeV: Code with Images for Faithful Visual Reasoning via Tool-Aware Policy Optimization

Xinhai Hou[1] Shaoyuan Xu[2] Manan Biyani[2] Moyan Li[2]
Jia (Kevin) Liu[3] Todd C Hollon[1] Bryan Wang[2]

[1] UNIVERSITY OF MICHIGAN
[2] amazon
[3] THE OHIO STATE UNIVERSITY

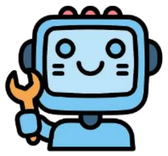
Problem shifts from single-step prediction to multi-step agentic tasks

We hope to leverage it for vision

Visual search as an agentic task (V*[1])

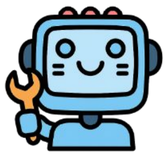
[1] Wu and Xie, CVPR 2024

Agent uses tools to answer the question



Seems to have three colors. Let me zoom in to further examine.

Cropped image shows a **flag** with 3 colors.

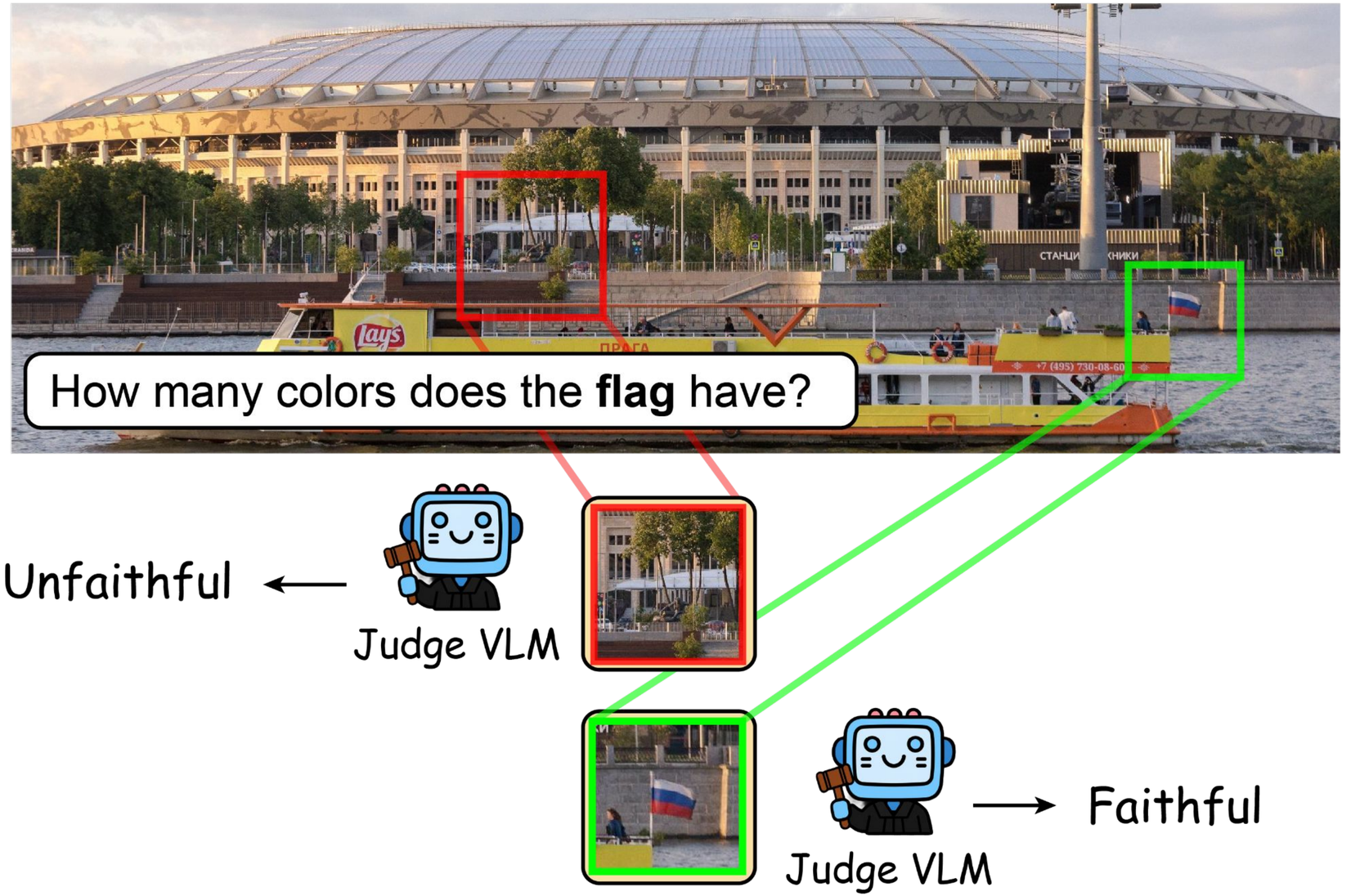


Seems to have three colors. Let me zoom in to confirm.

Cropped image shows a **flag** with 3 colors.

Correct answer with unfaithful tool use

Faithfulness Evaluation

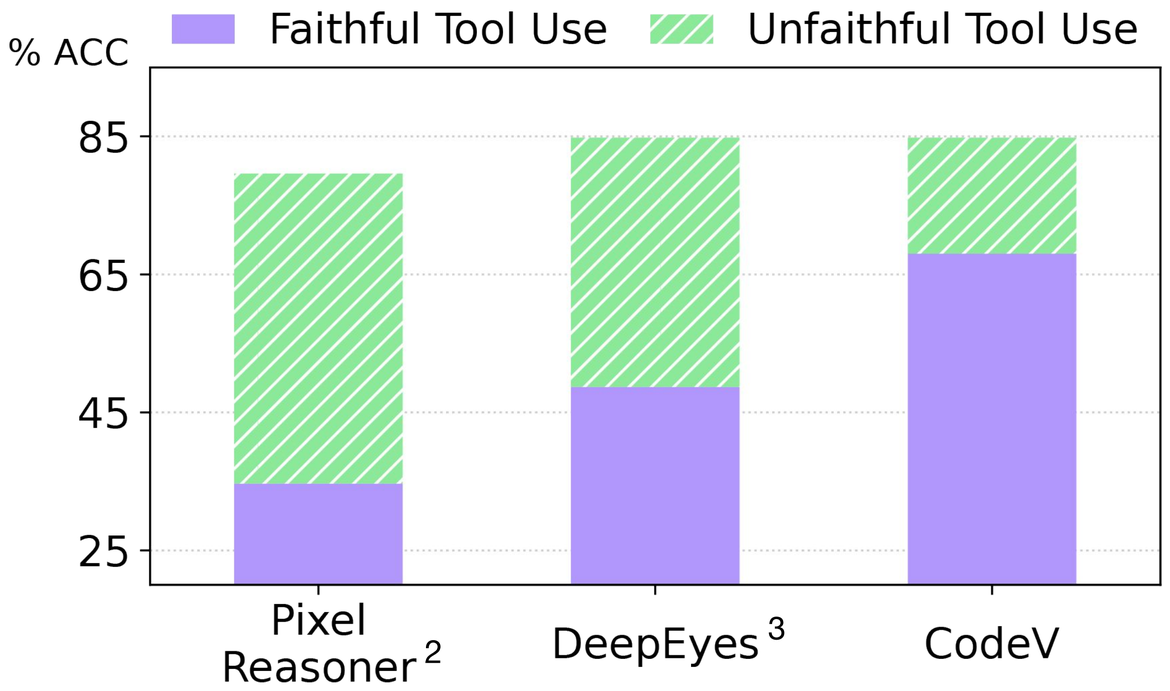


Faithfulness on V*[1]

Faithfulness for correct answer only.

Judge only checks tool outputs.

No evaluation on reasoning tokens.



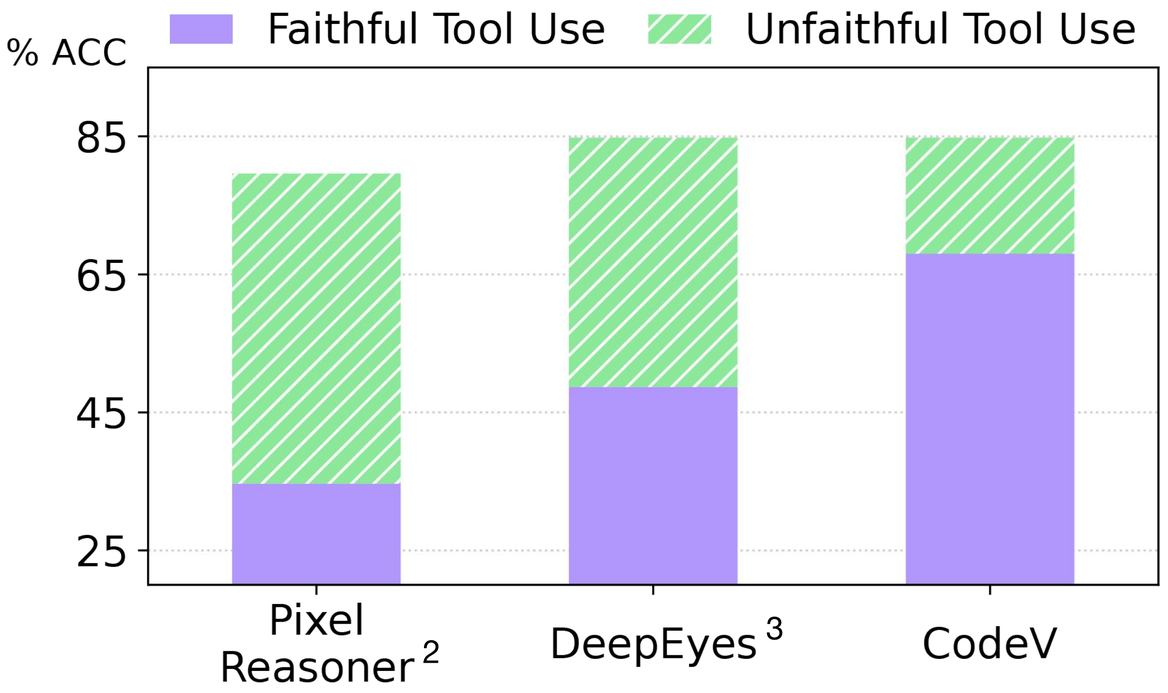
[1]Wu and Xie, CVPR 2024, [2] Su et al., NeurIPS 2025, [3] Zheng et al., ICLR 2026,

Faithfulness on V*[1]

Faithfulness for correct answer only.

Judge only checks tool outputs.

No evaluation on reasoning tokens.

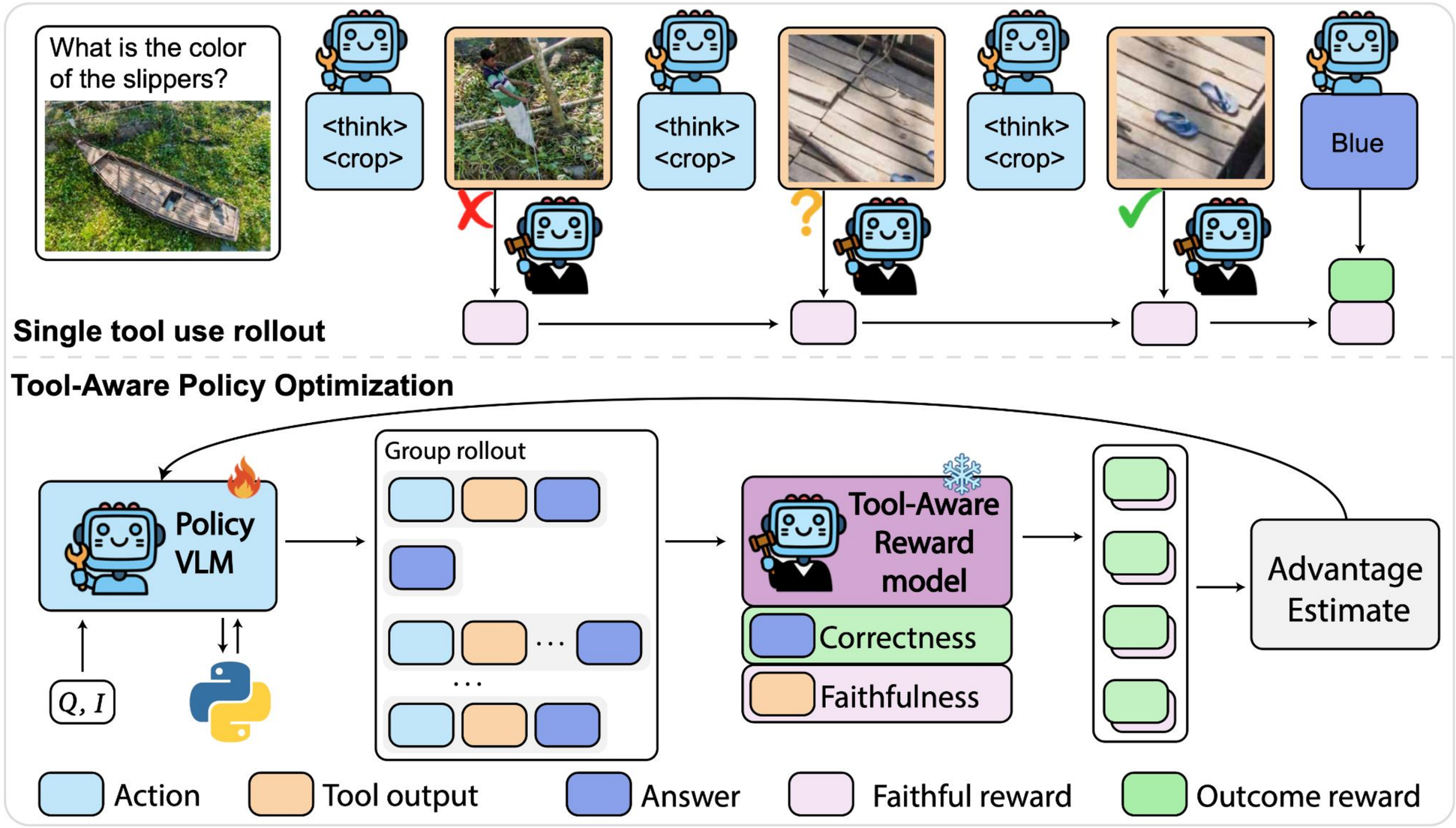


Hypothesis: Reward design is outcome dominant and sparse

[1]Wu and Xie, CVPR 2024, [2] Su et al., NeurIPS 2025, [3] Zheng et al., ICLR 2026,

How do we incentivize vision–language models to produce faithful reasoning grounded in tool outputs?

Tool-Aware Policy Optimization



Reward design

$$R(\tau) = \lambda_{\text{acc}}\, r^{\text{acc}}(\tau) + \lambda_{\text{tool}}\, r^{\text{tool}}(\tau)$$

$$r^{\text{tool}}(\tau) = \frac{1}{|\mathcal{T}_{\text{tool}}|} \sum_{t \in \mathcal{T}_{\text{tool}}} r_t^{\text{tool}}$$

Reward design

$$R(\tau) = \lambda_{\text{acc}}\, r^{\text{acc}}(\tau) + \lambda_{\text{tool}}\, r^{\text{tool}}(\tau)$$

$$r^{\text{tool}}(\tau) = \frac{1}{|\mathcal{T}_{\text{tool}}|} \sum_{t \in \mathcal{T}_{\text{tool}}} r_t^{\text{tool}}$$

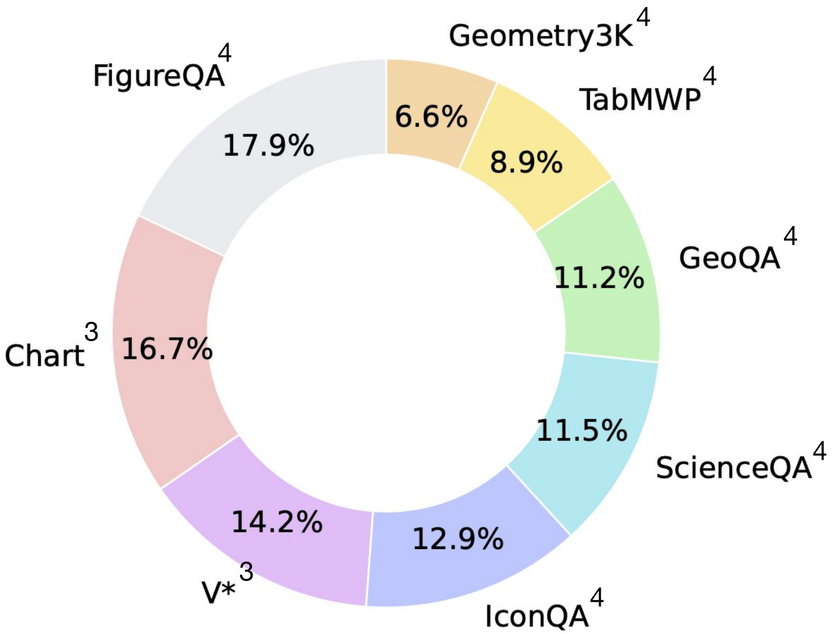
Tips:

Judge tool output instead of reasoning

$$\lambda_{\text{acc}} > \lambda_{\text{tool}}$$

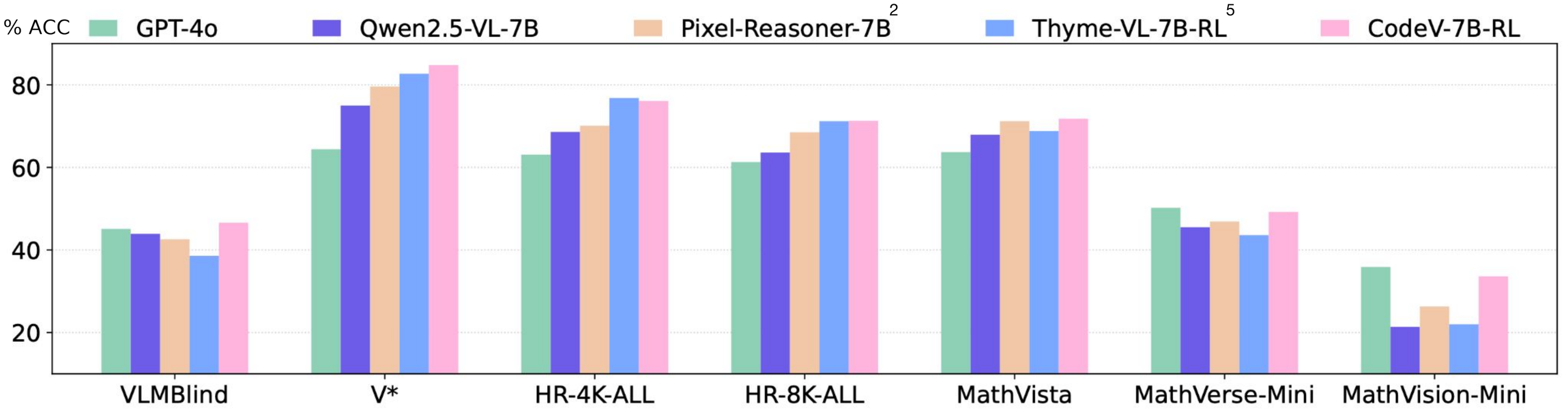
Experiment components

- SFT + RL two stage training
- RL environment with rubrics reward
- 8 training datasets
  - high resolution
  - visual search
  - figure/chart
  - geometry
- Evaluation set
  - perception
  - reasoning



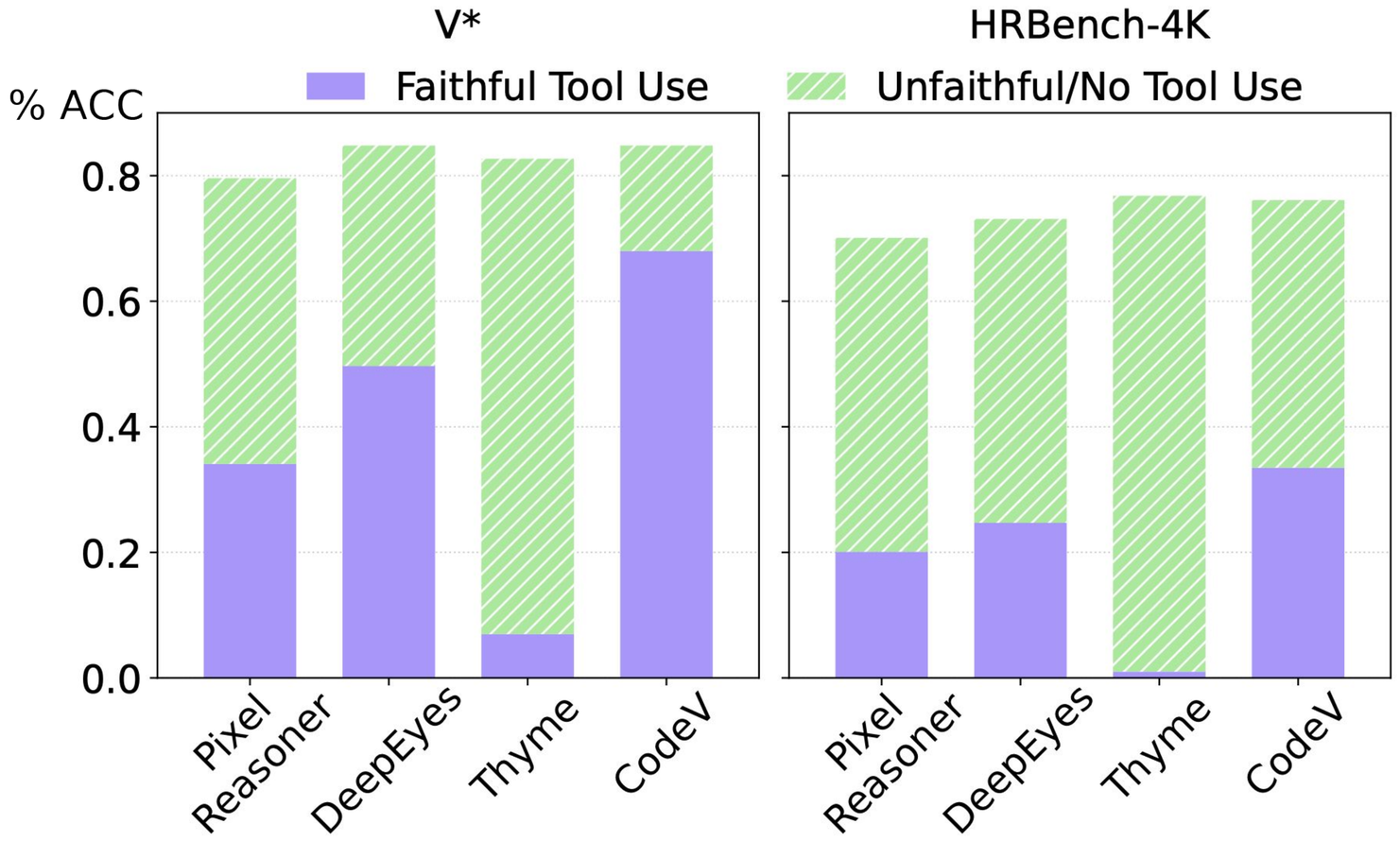
[3] Zheng et al., ICLR 2026, [4] Wang et al., NeurIPS 2025,

Main results

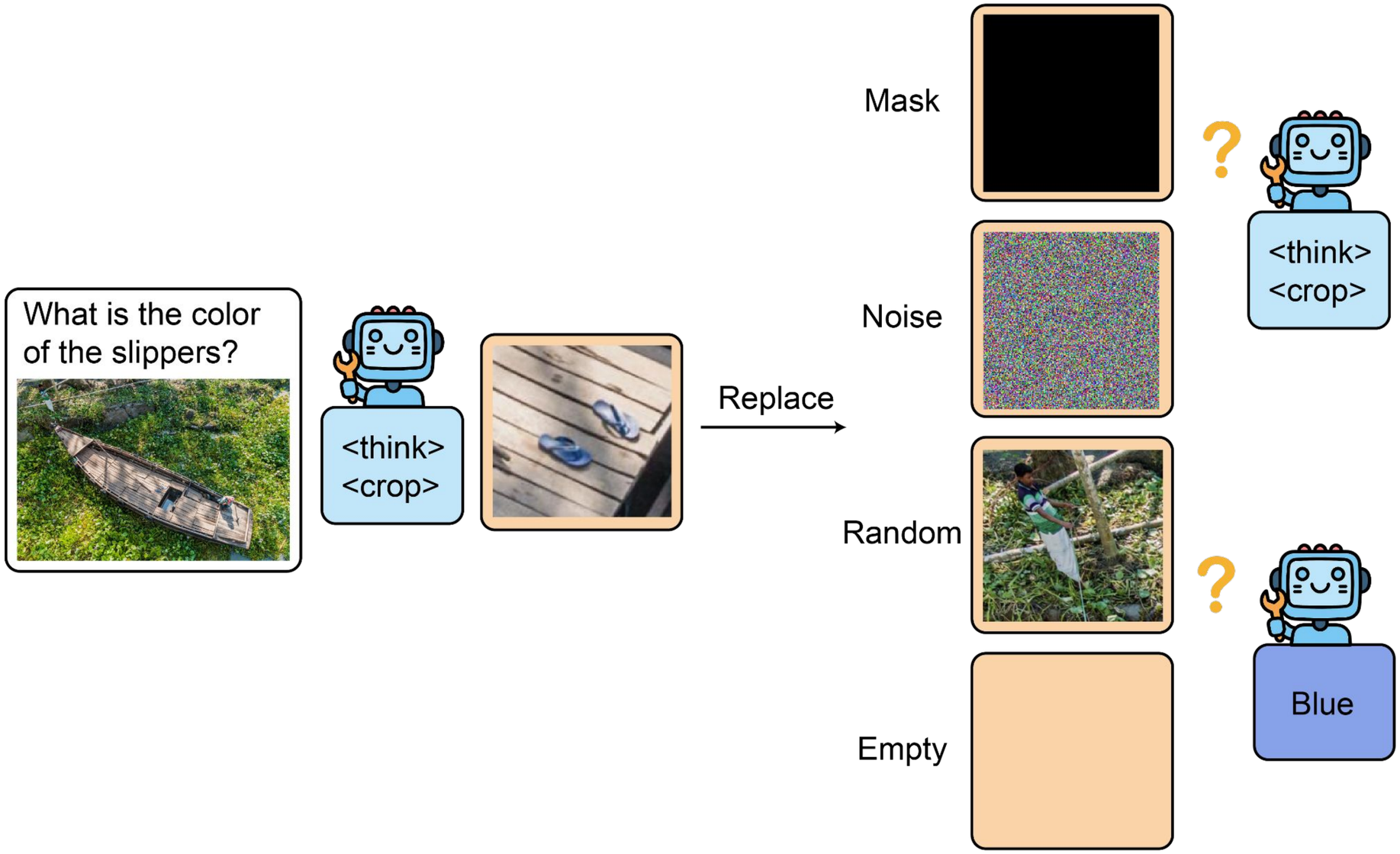


[2]Su et al., NeurIPS 2025, [5] Zhang et al., ICLR 2026

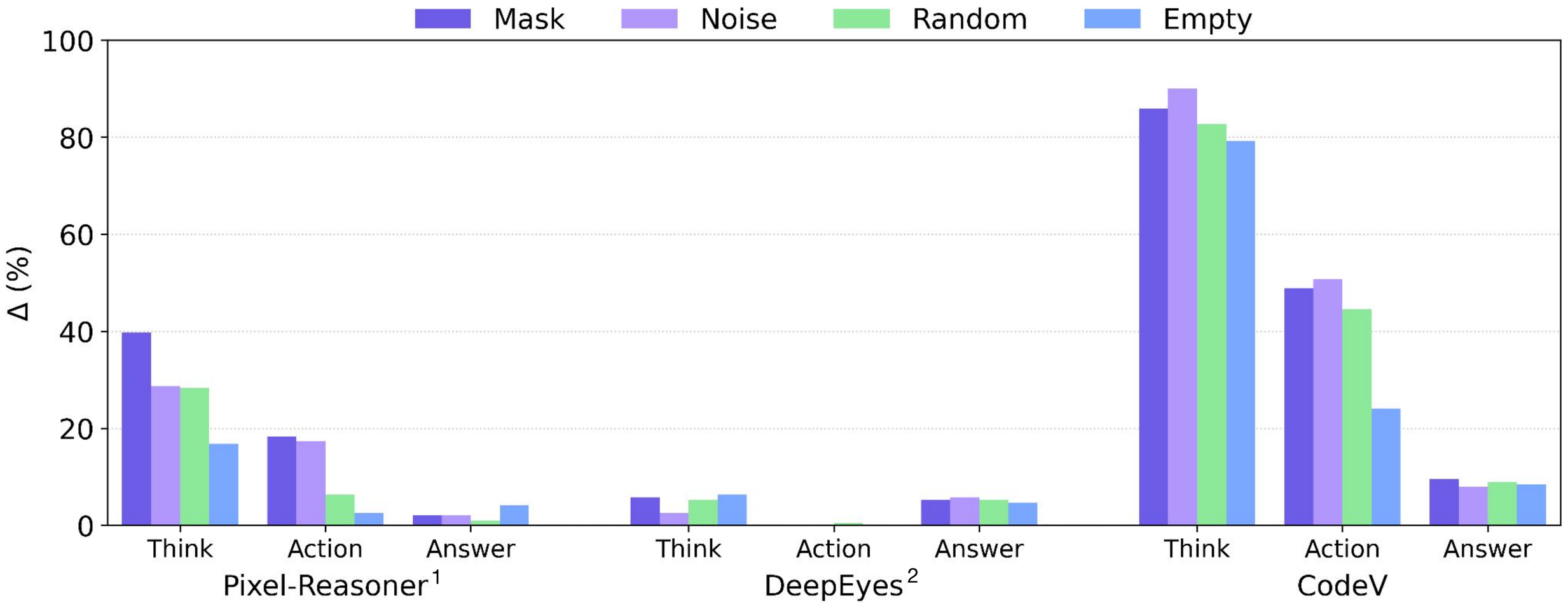
Faithfulness results



Perturbation study



Perturbation study



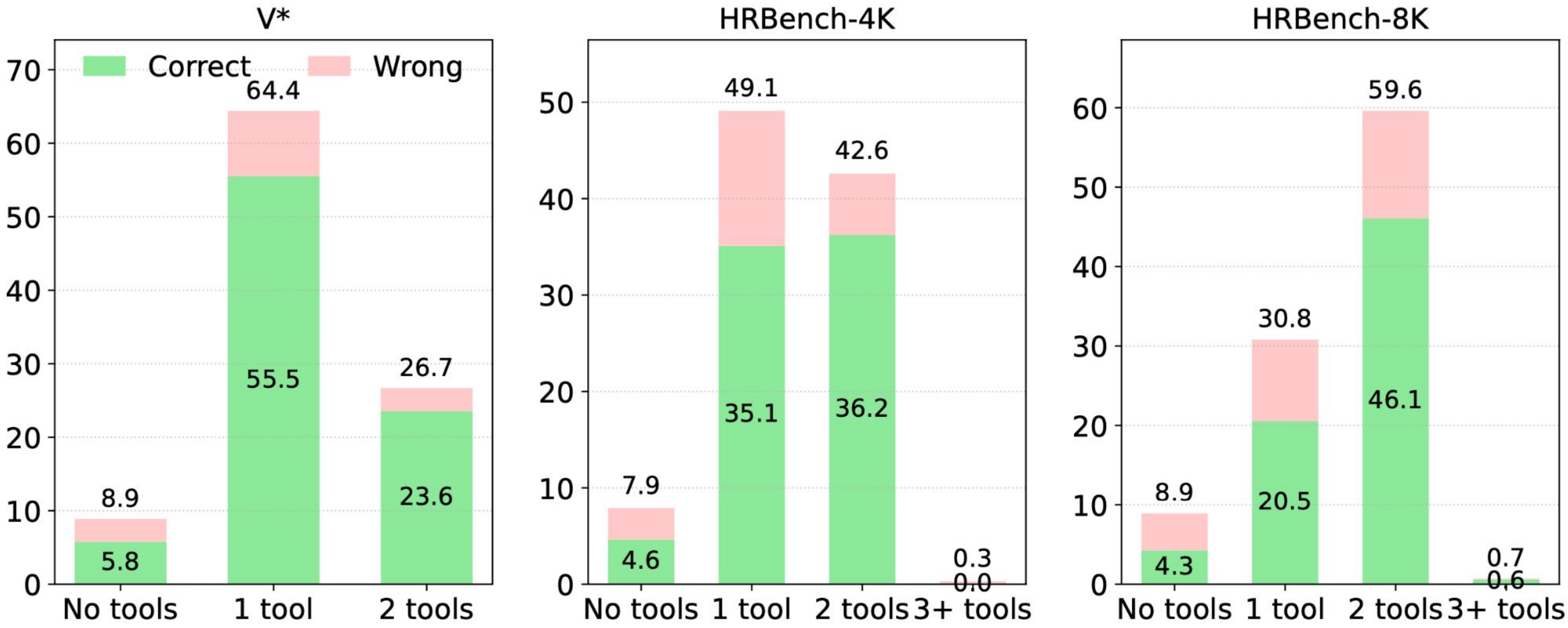
[1]Su et al., NeurIPS 2025, [2]Zheng et al., ICLR 2026

Ablation study

Model	VLMBlinds	V*	HRBench-4K			HRBench-8K		
			ALL	FSP	FCP	ALL	FSP	FCP
Training stage								
Qwen2.5-VL-7B	43.9	75.0	68.6	82.2	55.0	63.6	75.0	52.2
Zero-RL	46.6	78.5	73.0	89.0	57.0	69.9	84.2	55.5
Cold-start SFT	34.4	76.4	71.5	84.8	58.2	64.5	77.8	51.2
Reward design								
Accuracy only	43.8	82.7	74.9	89.8	60.0	68.8	82.8	54.8
+ Consistency	45.0	83.2	74.4	87.5	61.3	67.9	79.0	56.8
+ GPT-5-nano Judge	44.7	84.3	74.8	89.8	59.8	71.1	85.2	57.0
CodeV-7B-RL	**46.7**	**84.8**	**76.1**	**91.0**	**61.3**	**71.2**	**81.2**	**60.2**

Model	MathVista	CharXiv Reasoning	CharXiv Description	MMMU	MathVerse Mini	MathVision Mini
Training stage						
Qwen2.5-VL-7B	67.9	36.3	71.8	55.2	45.5	21.4
Zero-RL	69.8	38.2	71.5	**59.3**	48.2	29.9
Cold-start SFT	68.1	32.3	70.6	48.4	44.2	23.7
Reward design						
Accuracy only	69.0	39.0	71.4	52.8	47.9	27.3
+ Consistency	69.3	36.6	70.5	52.0	48.0	26.0
+ GPT-5-nano Judge	**72.4**	**39.6**	**72.5**	56.7	**49.7**	23.7
CodeV-7B-RL	71.8	39.3	72.0	**59.3**	49.2	**33.6**

Tool-use frequency



Summary and contributions

We reveal the unfaithful tool-use pattern in visual agent.

We propose TAPO, a process-level RL framework, emphasize a rubric-reward system.

CodeV achieves empirical gains in faithfulness and accuracy.

We release the open recipe of the RL training, RL environment and evaluation.

Limitations and future work

Static rubrics reward is expensive and can still be hacked.

Verifiable signal can be hard to find for general tasks.

RL algorithm only explore the GRPO, we expect more variants like DPO, PPO.

Reference list

[1] Penghao Wu and Saining Xie. V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13084–13094, 2024.

[2] Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel Reasoner: Incentivizing Pixel Space Reasoning via Curiosity-Driven Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[3] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing “Thinking with Images” via Reinforcement Learning. In Proceedings of the Fourteenth International Conference on Learning Representations, 2026.

[4] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. SoTA with Less: MCTS-Guided Sample Selection for Data-Efficient Visual Reasoning Self-Improvement. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[5] Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think Beyond Images. In Proceedings of the Fourteenth International Conference on Learning Representations, 2026.


LOVE


Haidilao Hot Pot
海底捞火锅
THE NORTH FACE

Q&A

Code

Poster #7
16:00-18:00
June 6