Vero: An Open RL Recipe for General Visual Reasoning

Gabriel Sarch* Linrong Cai* Qunzhong Wang Haoyang Wu Danqi Chen Zhuang Liu†
Princeton University
* Project Leads † Corresponding Author

Models Data Code Project Page

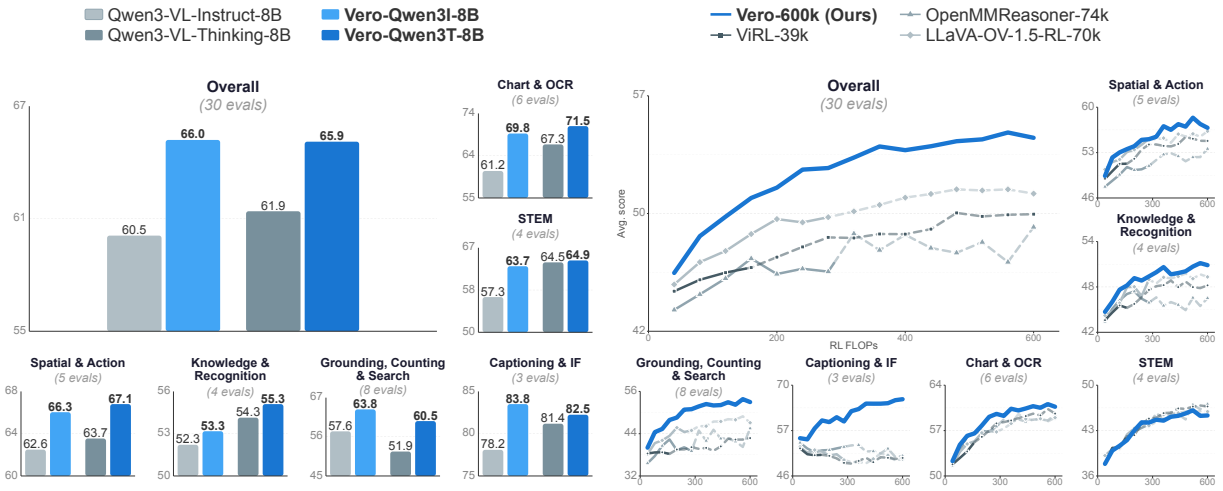

Figure 1. **Vero** improves performance across the overall benchmark and six task categories. **Left.** Model comparison across overall and category scores. **Right.** Scaling trends versus RL FLOPs. All runs are finetuned from Qwen2.5-VL-7B-Instruct.

Abstract

*The strongest vision-language models (VLMs) rely on proprietary reinforcement learning (RL) pipelines, while broad multi-task RL remains difficult because heterogeneous visual problems transfer weakly across tasks. We introduce **Vero**, a family of fully-open VLMs trained with a carefully curated collection of 600K RL samples from 59 datasets spanning six core task categories. **Vero** achieves state-of-the-art performance across a wide range of visual reasoning tasks, improving over four base models by 3.6–5.5 points on average across 30 challenging benchmarks spanning the six core task categories. Starting from Qwen3-VL-8B-Instruct, **Vero** outperforms Qwen3-VL-8B-Thinking on 23 of 30 benchmarks without using any additional proprietary thinking data. On MiMo-VL-SFT, **Vero** surpasses MiMo-VL-RL, which relies on a proprietary RL recipe. Systematic ablations reveal that different task categories elicit qualitatively distinct reasoning patterns that transfer poorly in isolation, suggesting that broad data coverage is the primary driver of strong RL scaling. All data, code, and models are released.*

1. Introduction

Training models to reason through explicit chain-of-thought (CoT) has become a powerful paradigm for improving large language models (LLMs) and vision-language models, with on-policy reinforcement learning as a key driver [5, 12, 13]. Yet the strongest visual reasoning models rely on proprietary RL pipelines with non-public data, while fully open efforts such as OpenMMReasoner [21] and VL-Rethinker [15] focus primarily on visual math. As we show in Figure 1 and Section 4, training on a single task category does not generalize to other visual capabilities.

We introduce **Vero**, a family of fully open VLMs trained with a single-stage RL recipe on top of existing VLMs. We curate 600K high-quality samples from 59 datasets spanning six core task categories and pair them with task-routed reward functions, without warm start, staged RL, or proprietary data. Through systematic ablations, we find that data diversity is the critical ingredient: different task categories elicit distinct reasoning patterns that transfer poorly in isolation. Our best Qwen3-based variants achieve 66.0 and 65.9 overall average, the highest among the 8B VLMs we eval-

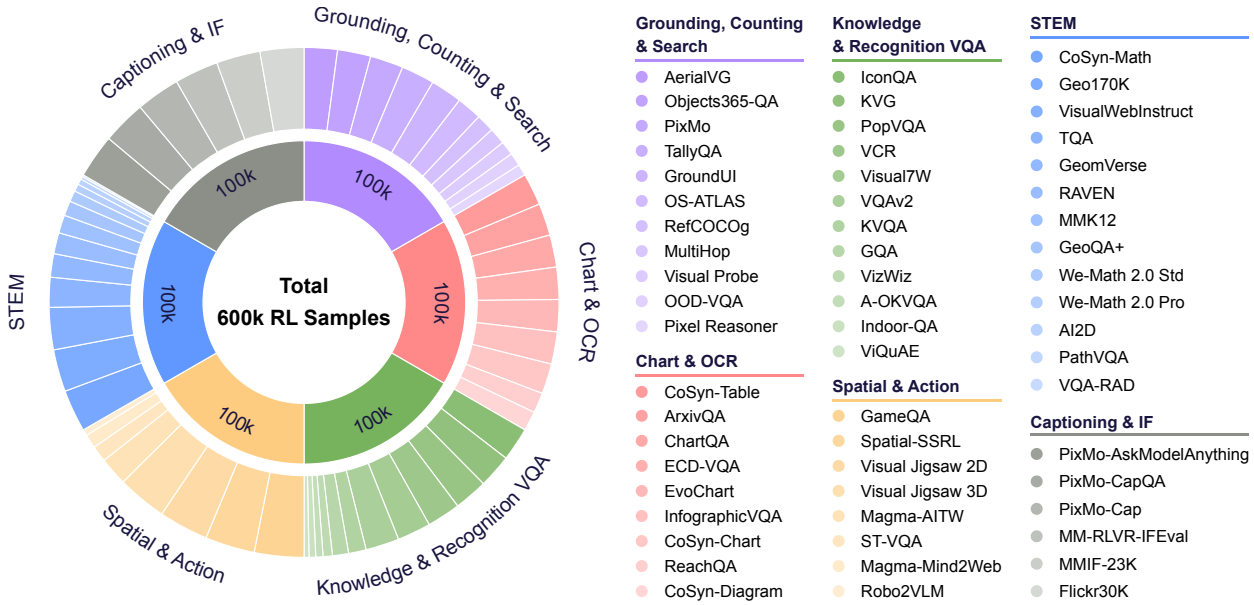

Figure 2. Composition of the RL training data. The inner ring shows six task categories, and the outer ring shows their 59 datasets.

	Chart& OCR	STEM	Spatial& Action	Knowl.& Recog.	Ground.& Count.	Bench. Avg.
equal ratios	**+8.6**	+6.2	**+5.6**	+1.8	+5.6	**+5.8**
ratio $\propto (1-\text{acc.})^{\alpha}$	+6.8	**+6.5**	+4.3	**+2.4**	+5.2	+5.2
ratio $\propto \text{area}^{\alpha}$	+7.0	+5.3	+4.1	+1.4	**+6.2**	+5.2
ratio $\propto \text{length}^{\alpha}$	+7.5	+6.4	+4.5	+1.7	+3.8	+4.8
w/o Knowl. & Recog.	+6.4	**+6.5**	+4.8	+1.9	+4.7	+4.9

Table 1. Performance of different task category weightings schemes. Values are score changes (Δ) over the base model.

uate. Starting from Qwen3-VL-8B-Instruct, **Vero** outperforms Qwen3-VL-8B-Thinking on 23 of 30 benchmarks, and applied on top of Qwen3-VL-8B-Thinking it improves this further to 24 of 30. We also surpass MiMo-VL-7B-RL from the same base checkpoint using our fully open recipe. All data, code, and models are released.

2. Vero

2.1. Task Definitions

We train a VLM π_θ via RL to maximize expected reward across diverse visual reasoning tasks. Given visual input v and text query q, the model generates $y \sim \pi_\theta(\cdot \mid v, q)$, evaluated by $R(y, y^*)$ against ground truth y^*.

Task taxonomy. We organize training data into six categories (Figure 2), motivated by the finding that training on any single category fails to transfer reliably to others (Section 4): **STEM** (13 datasets), math, science, and medical image reasoning; **Spatial & Action** (8), embodied reasoning, UI navigation, and 3D understanding; **Knowledge & Recognition** (12), VQA combining recognition with external knowledge; **Chart & OCR** (9), reasoning over documents, charts, and infographics; **Grounding, Counting & Visual Search** (11), object localization, counting, and visual search; **Captioning & Instruction Following** (6), open-ended description and instruction adherence.

2.2. Dataset Curation, Filtering, and Mixtures

Sourcing and filtering. We curate 600K samples from 59 datasets. Starting from over 250 candidates, we apply heuristic filters (minimum 1K examples, adequate resolution, no binary-only answers) and manual inspection of $\sim$50 examples per dataset for correctness, unambiguity, and verifiability, retaining 59 datasets. We apply model-based question filtering using Qwen3-VL-235B to remove ambiguous or unverifiable questions, and answer canonicalization to normalize ground-truth formats.

Data mixtures. We compare four task category weighting schemes (uniform, difficulty-weighted, dataset-size-weighted, reasoning-length-weighted). Uniform sampling achieves the highest overall gain (+5.8 pts), as alternatives improve individual categories at the cost of others (Table 1).

2.3. Reinforcement Learning

Algorithm. We use GSPO [22], which computes a sequence-level importance ratio, with asymmetric clipping [18], no KL penalty, and a soft overlong penalty.

	Vero (Ours)				Open Weights Models				Fully Open RL Recipes			Propr.
	Vero Qwen3I-8B	**Vero Qwen3T-8B**	**Vero Qwen25-7B**	**Vero MiMo-7B**	**Q3VL 8B-Ins**	**Q3VL 8B-Thk**	**Q25VL 7B-Ins**	**MiMoVL 7B-RL**	**LLaVA OV1.5-RL**	**VL-Re thinker**	**Mo2-O 7B**	**GPT-5 Nano**
RL Initial Model	Qwen3VL 8B Inst	Qwen3VL 8B Think	Qwen25VL 7B Inst	MiMoVL 7B-SFT	N/A	N/A	N/A	MiMoVL 7B-SFT	LLaVA OV1.5-Ins	Q25VL 7B-Ins	SFT Only	N/A
Chart & OCR												
ChartQA-Pro	60.2 +15.9	**62.9** +4.0	52.0 +8.7	**62.9** +11.8	44.3†	58.9†	43.3†	61.1†	-	-	31.4†	56.0†
ChartQA	91.6 +2.0	90.8 +2.2	90.0 +2.7	90.4 +0.3	89.6	88.6	87.3	**94.4**	87.4	-	75.2†	80.1†
InfoVQA	87.8 +4.7	88.2 +2.2	81.9 -0.7	87.0 +5.1	83.1	86.0	82.6	**90.1**	76.6	-	60.3†	67.9†
CharXiv$_{Reason}$	53.7 +7.3	59.9 +6.9	44.6 +2.1	**66.7** +13.6	46.4	53.0	42.5	60.9†	-	-	35.6†	51.2†
ChartMuseum	49.6 +9.6	**51.2** +6.8	36.0 +9.2	47.8 +7.6	40.0	44.4	26.8	48.7†	-	-	30.3†	48.0†
EvoChart	75.7 +11.7	**75.9** +2.9	66.9 +4.1	73.9 +2.9	64.0†	73.0†	62.8†	73.4†	-	-	51.0†	63.3†
Category Avg	69.8 +8.5	**71.5** +4.2	61.9 +4.4	71.5 +6.9	61.2	67.3	57.6	71.4	-	-	47.3	61.1
STEM												
MMMU-Pro$_{Std}$	59.8 +3.9	59.5 -0.9	43.4 +5.1	58.8 +2.7	55.9	60.4	38.3	59.4†	39.9	41.7	31.9†	**61.3**†
MMMU-Pro$_{Vis}$	57.2 +15.1	**57.5** +4.1	39.6 +7.3	56.0 +8.3	42.1†	53.4†	32.3†	49.8†	35.7	-	16.0†	53.1†
MathVision	59.0 +5.1	**63.5** +0.8	29.1 +4.0	57.4 +0.6	53.9	62.7	25.1	58.8†	34.4	-	21.3†	61.7†
MathVista$_{testmini}$	78.7 +1.5	79.2 -2.2	74.5 +6.3	79.0 -1.6	77.2	**81.4**	68.2	80.4†	72.3	73.7	53.6†	70.2†
Category Avg	63.7 +6.4	**64.9** +0.5	46.6 +5.7	62.8 +2.5	57.3	64.5	41.0	62.1	45.6	-	30.7	61.6
Spatial & Action												
Blink	68.7 -0.4	66.3 +1.6	57.5 +1.1	61.2 -1.2	**69.1**	64.7	56.4	64.5†	-	-	56.4†	59.3†
ERQA	43.2 -2.6	**47.2** +0.4	40.0 -1.8	39.8 +0.3	45.8	46.8	41.8†	43.5†	-	-	43.5†	45.5†
GameQA$_{Lite}$	52.3 +18.3	**54.9** +15.1	46.7 +20.6	53.2 +7.0	34.0†	39.8†	26.1†	49.8†	-	-	29.6†	45.9†
EmbSpatial	79.2 +0.7	79.8 -1.3	72.0 +1.3	70.0 -2.5	78.5	**81.1**	70.7†	70.2†	-	-	68.1†	74.2†
CV Bench	**87.9** +2.4	87.4 +1.4	81.1 +0.7	83.6 +1.0	85.5†	86.0†	80.4†	83.5†	82.9	-	81.7†	82.5†
Category Avg	66.3 +3.7	**67.1** +3.4	59.5 +4.4	61.6 +0.9	62.6	63.7	55.1	62.3	-	-	55.9	61.5
Knowledge & Recognition												
RealWorldQA	73.3 +1.8	71.5 -2.0	71.6 +3.1	69.3 +1.7	71.5	**73.5**	68.5	68.6†	68.4	-	73.3†	65.9†
SimpleVQA$_{En}$	45.2 +1.0	46.2 +1.3	**50.0** +5.5	47.5 +5.6	44.2†	44.9†	44.5†	40.9†	-	-	30.8†	36.6†
FVQA	24.6 -1.4	22.0 -2.2	26.3 +4.4	30.1 +0.3	26.0	24.2	21.9	**31.8**†	-	-	17.7†	29.4†
MM-Vet v2	70.2 +2.6	**81.6** +7.1	66.3 +3.4	79.2 +19.5	67.6	74.5	62.9	61.2†	-	-	60.8†	71.5†
Category Avg	53.3 +1.0	55.3 +1.1	53.5 +4.1	**56.5** +6.8	52.3	54.3	49.5	50.6	-	-	45.6	50.9
Grounding, Counting, Search												
CountBenchQA	90.4 +1.6	**93.1** +3.5	84.9 -1.0	85.3 +0.0	88.8	89.6	85.9†	86.4†	86.8	-	89.4†	75.4†
CountQA	33.9 +5.4	**37.6** +5.8	24.1 +3.2	25.1 -0.5	28.5†	31.8†	20.9†	27.4†	-	-	32.1†	25.7†
MMERealWorld	**57.8** +10.7	51.5 +6.7	52.4 +7.9	52.1 +8.3	47.1	44.8	44.5	48.7†	-	-	44.4†	49.8†
VStarBench	**89.5** +7.3	84.3 +7.9	86.4 +6.3	82.2 +0.0	82.2	76.4†	80.1†	84.3	79.1	-	73.8†	71.2†
AerialVG	30.0 -2.2	31.0 +18.4	27.4 +5.1	19.6 -0.5	**32.2**†	12.6†	22.3†	22.2†	-	-	-	-
VisualProbe	**53.9** +6.2	47.3 +8.0	48.8 +2.8	50.4 +3.6	47.7†	39.3†	46.0†	52.2†	-	-	34.8†	41.5†
ScreenSpot	**93.6** +7.0	92.0 +6.5	89.9 +4.3	89.2 +1.3	86.6†	85.5†	85.6†	87.3†	-	-	75.8†	-
ScreenSpotPro	**61.4** +13.6	47.6 +12.1	38.1 +14.2	34.5 +1.7	47.8	35.5	23.9†	37.4†	-	-	19.3†	-
Category Avg	**63.8** +6.2	60.5 +8.6	56.5 +5.4	54.8 +1.7	57.6	51.9	51.1	55.7	-	-	-	-
Captioning & IF												
MM-MTBench	**80.3** +5.9	74.3 -3.5	66.1 +7.2	75.8 +0.1	74.4	77.8	58.9	79.2†	-	-	33.3†	72.7†
MIABench	93.5 +2.4	**93.7** +2.2	82.8 +1.0	91.5 +3.1	91.1	91.5	81.8	88.4†	-	-	77.9†	92.0†
MMIFEval	77.7 +8.5	**79.6** +4.6	62.8 +9.2	78.3 +9.1	69.2	75.0	53.6	66.1	-	-	54.4†	78.0†
Category Avg	**83.8** +5.6	82.5 +1.1	70.6 +5.8	81.9 +4.1	78.2	81.4	64.8	77.9	-	-	55.2	80.9
Overall Avg	**66.0** +5.5	65.9 +4.0	57.8 +4.9	63.3 +3.6	60.5	61.9	52.9	62.4	-	-	-	-

Table 2. **Vero achieves state-of-the-art performance across multiple task categories.** Vero columns show the initial models trained with RL on our dataset; +x / -x deltas indicate improvement/decline over the respective initial model. † indicates results evaluated by us. All other results are taken from official technical reports.

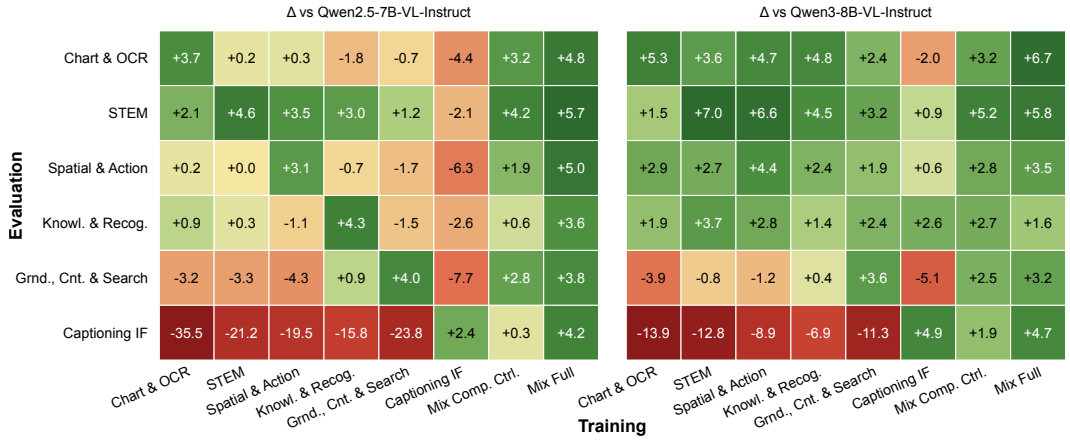

Figure 3. Cross-task generalization. Each row shows a model trained on a single task category or mixture. Values are absolute score changes relative to the base model.

Reward formulation. We use ten task-routed verifiers: *string match*, *multiple choice*, *numeric* (via `math_verify` [7]), *list match*, *ordering* [16], *web action* [11], *grounding* (IoU/F1) [17], *clicking* [8], *instruction following* [6, 10], and *LLM-as-judge* [9] (Qwen3-32B). A format reward additionally checks for `<think>...</think><answer>...</answer>` structure. We show in Section 5 that this design outperforms simpler alternatives.

3. Experiments

Evaluation suite. We evaluate on **VeroEval**, 30 benchmarks spanning our six categories, using the lmms-eval [20] framework with official evaluation protocols.

Baselines. We compare against base VLMs (Qwen2.5-VL-7B-Instruct [3], Qwen3-VL-8B-Instruct/Thinking [2], Molmo2-O-7B [4]), open RL models (VL-Rethinker-7B [15], LLaVA-OV-1.5-RL [1]), proprietary RL (MiMo-VL-7B-RL [19]), and gpt-5-nano [14].

3.1. Evaluation Results

Results are in Table 2. Our best models **Vero-Qwen3I-8B** and **Vero-Qwen3T-8B** achieve 66.0 and 65.9 overall average, the highest among all 8B VLMs we evaluate. Training yields consistent improvements across four base models (+3.6 to +5.5 pts averaged over 30 benchmarks).

Vero-Qwen3I-8B outperforms Qwen3-VL-8B-Thinking on 23 of 30 benchmarks, while **Vero-Qwen3T-8B** improves this to 24 of 30. The largest category-level gains over Qwen3-VL-8B-Thinking are on Grounding, Counting & Search (+11.9 for **Vero-Qwen3I-8B**) and Chart & OCR (+4.2 for **Vero-Qwen3T-8B**). **Vero-Qwen25-7B** surpasses the substantially stronger Qwen3-VL-8B-Instruct on Chart & OCR and Knowledge & Recognition despite a 7.6-point base model gap. **Vero-MiMo-7B** outperforms MiMo-VL-7B-RL on 4 of 6 category averages, showing that a fully open recipe can surpass proprietary pipelines from the same checkpoint.

4. Cross-Task Generalization

We train models on each individual task category (100k samples) and evaluate across all six (Figure 3). Single-task training frequently produces negative transfer: on Qwen2.5-VL, nearly all single-category models degrade Grounding performance, and training on any non-captioning category severely degrades Captioning & IF (-15.8 to -35.5). In contrast, diverse task mixing eliminates negative transfer, achieving positive gains across all categories on both Qwen2.5-VL and Qwen3-VL base models.

5. Ablations

All ablations use Qwen2.5-VL-7B-Instruct trained for 1 epoch on the 600k mixture.

Filtering improves data quality. Question filtering and answer canonicalization improve data quality, with the largest gains on Spatial & Action (+1.9) and Knowledge & Recognition (+2.1). Our curated data also outperforms alternative dataset sources on Chart & OCR and STEM.

Reward design matters for multi-task training. Our multi-route reward design reaches 57.2 overall average, outperforming math_verify [7] (51.8).

RL outperforms SFT. On the same data, RL reaches 57.2 overall versus 52.8 for SFT and improves every category. SFT on our data also exceeds a strong alternative SFT data source (46.2), confirming the value of our curation. **Conclusion.** We presented **Vero**, a fully open VLM reasoning family trained with single-stage RL on 600K samples from 59 datasets spanning six capability categories. Our best model achieves state-of-the-art performance among 8B VLMs and outperforms Qwen3-VL-8B-Thinking on up to 24 of 30 benchmarks.

References

[1] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] 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 preprint arXiv:2502.13923*, 2025.

[4] Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026.

[5] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*, 645, 2025.

[6] Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Mm-ifengine: Towards multimodal instruction following. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1099–1109, 2025.

[7] Hynek Kydlíček. Math-verify: Math verification library. https://github.com/huggingface/Math-Verify, 2025.

[8] Zhengxi Lu, Yuxiang Chai, Yaxuan Guo, Xi Yin, Liang Liu, Hao Wang, Han Xiao, Shuai Ren, Guanjing Xiong, and Hongsheng Li. Ui-r1: Enhancing efficient action prediction of gui agents by reinforcement learning. *arXiv preprint arXiv:2503.21620*, 2025.

[9] OLMo Team, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

[10] Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[11] Gabriel Herbert Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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

[13] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[14] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[15] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-Rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. In *NeurIPS*, 2025.

[16] Penghao Wu, Yushan Zhang, Haiwen Diao, Bo Li, Lewei Lu, and Ziwei Liu. Visual jigsaw post-training improves mllms, 2025.

[17] En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, et al. Perception-R1: Pioneering perception policy with reinforcement learning. In *NeurIPS*, 2025.

[18] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *NeurIPS*, 2025.

[19] Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, et al. MiMo-VL technical report. *arXiv preprint arXiv:2506.03569*, 2025.

[20] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, 2025.

[21] Kaichen Zhang, Keming Wu, Zuhao Yang, Bo Li, Kairui Hu, Bin Wang, Ziwei Liu, Xingxuan Li, and Lidong Bing. Openmmreasoner: Pushing the frontiers for multimodal reasoning with an open and general recipe. *arXiv preprint arXiv:2511.16334*, 2025.

[22] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.