Vero: Open Reinforcement Learning Recipes for Visual Reasoning

Gabriel Sarch*, Linrong Cai*, Qunzhong Wang, Haoyang Wu
Danqi Chen, and Zhuang Liu†

Princeton University, Princeton, NJ, USA

Models Data Code Project Page

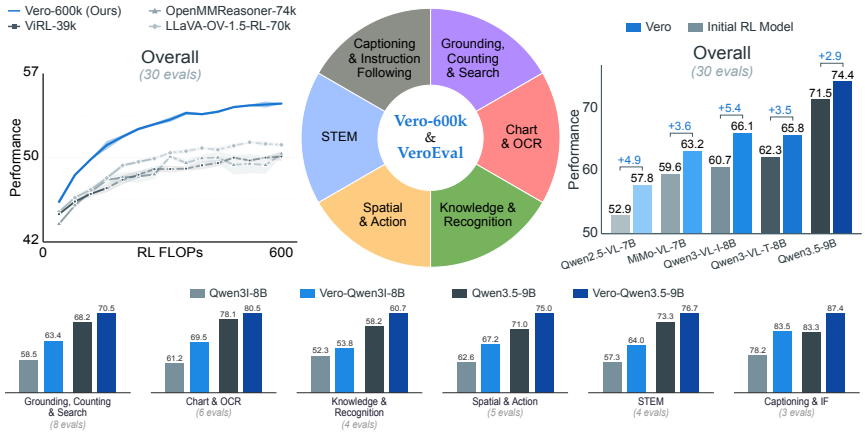

Fig. 1: Vero achieves state-of-the-art performance across six task categories using a fully open RL recipe. Top left: Training curves versus existing open RL datasets (Qwen2.5 -VL-7B-Instruct; dashed = beyond one epoch; mean ± sem, 3 seeds). **Top center:** Task categories in **Vero-600K**/**VeroEval**. **Top right:** **Vero** (Avg@5) vs. each initial RL model on 30 **VeroEval** benchmarks. **Bottom:** Per-category top right for **Vero** on Qwen3-VL-8B-Instruct and Qwen3.5-9B.

Abstract. What does it take to build a visual reasoner that works across charts, science, spatial understanding, and open-ended tasks? The strongest vision-language models (VLMs) suggest that broad visual reasoning is within reach, yet their closed data and reinforcement learning (RL) pipelines make their gains difficult to study, reproduce, or extend. We introduce **Vero**, a family of *fully open* VLMs that match or exceed existing open-weight models across diverse visual reasoning tasks. We scale RL data and rewards across six broad task categories, constructing **Vero-600K**, a 600K-sample dataset from 59 datasets, and designing task-routed rewards that handle heterogeneous answers. Across **VeroEval**, our 30-benchmark suite, **Vero-600K** outperforms existing RL datasets under controlled comparisons. Applied to five starting models, **Vero** variants gain 2.9–5.4 points on average over their initial models. Notably, **Vero-Qwen3I-8B**, trained on the Instruct model, surpasses Qwen3-VL-8B-Thinking by 3.8 points on average without additional distillation. Systematic ablations reveal that different task categories elicit distinct reasoning patterns and that broad gains depend on learning them jointly rather than in isolation. All data, code, and models are released.

Keywords: vision-language models · reinforcement learning · reasoning

*Project lead. †Corresponding author.

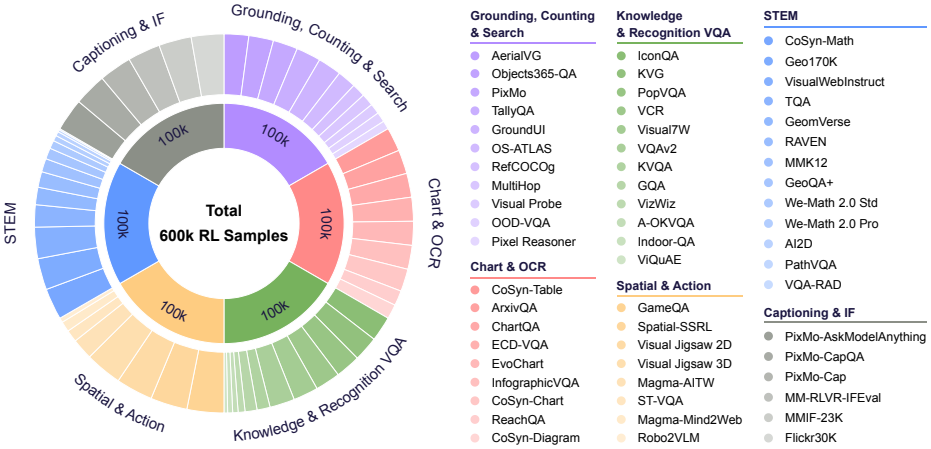

Fig. 2: Composition of Vero-600K. The inner ring shows six task categories, each allocated 100K samples (600K total), and the outer ring shows their 59 constituent datasets. The categories represent real-world use cases and cover distinct visual reasoning capabilities (Sections 5 and 6). Categories are uniformly sampled to balance learning across tasks.

1 Introduction

Vision-language models (VLMs) are increasingly expected to reason across a wide range of visual tasks, from chart and scientific interpretation to spatial understanding and open-ended questions. Reinforcement learning (RL) has emerged as a key driver of this progress, with methods such as PPO [71] and GRPO [74] enabling models to learn from their own generations through reward signals. Recent multimodal models such as GPT-5 [77], Qwen3-VL [3], and Kimi K2.5 [38] demonstrate that RL drives substantial improvement in multimodal reasoning.

Yet the strongest existing visual reasoning models are products of proprietary RL pipelines with non-public data and undisclosed reward designs. Models such as Qwen3-VL [3] release weights and are widely adopted, but do not release RL training code or datasets. Accompanying technical reports often omit detailed ablations of design choices, making it difficult to systematically study what drives performance. Meanwhile, fully open efforts such as OpenMMReasoner [110] and VL-Rethinker [87] focus primarily on visual math, covering only a narrow subset of visual tasks. However, as we show in Sections 5 and 6, training on a single task category does not generalize to other visual capabilities, in both task performance and chain-of-thought behavior. More broadly, applying RL across heterogeneous visual reasoning tasks is challenging, as diverse task mixtures induce interference, weak transfer, and optimization imbalance unless the training distribution and rewards are carefully designed [27, 70, 84]. This leaves a central question: *what does it take to train a broadly capable visual reasoner?*

We show that a single-stage RL recipe with diverse and high-quality data suffices. We introduce **Vero**, a family of fully open VLMs trained with RL on

top of existing models to perform strongly across diverse visual tasks. Our recipe centers on careful dataset selection, sample filtering, and task-balanced mixing: we build **Vero-600K**, a 600K-sample training set from 59 datasets spanning six core task categories (Chart & OCR; STEM; Spatial & Action; Knowledge & Recognition; Grounding, Counting & Search; and Captioning & Instruction Following), and pair it with task-routed reward functions. No additional warm start, no staged RL, and no proprietary data. Alongside the training data, we assemble **VeroEval**, a comprehensive evaluation suite of 30 benchmarks spanning all six categories. Figure 2 summarizes the composition of **Vero-600K**.

Through systematic ablations of dataset selection, sample filtering, mixture strategies, and reward design, we find that data diversity is the critical ingredient. Different task categories elicit qualitatively distinct reasoning patterns that transfer poorly in isolation: for example, STEM tasks trigger elevated backtracking while grounding tasks suppress introspective behaviors in favor of directed visual search. Producing a generally capable model therefore requires broad task coverage, so that the model learns these distinct reasoning patterns jointly rather than in isolation. We additionally find that (1) uniform mixture weighting across task categories outperforms schemes based on accuracy, reasoning length, or image size; (2) multi-task training necessitates an expressive reward design; and (3) including open-ended tasks is necessary to preserve visual chat ability.

Vero achieves strong overall performance across model families (Figure 1). Training on six different initial models yields consistent improvements, with gains of +2.9 to +5.4 points over five post-trained models, averaged over 30 benchmarks. Applied directly to the pretrained-only Qwen3.5-9B-Base, our recipe yields +12.9 points, reaching 73.0 overall, second only to **Vero-Qwen35-9B**, without any distillation warm start. **Vero-Qwen35-9B**, trained from Qwen3.5-9B, reaches 74.4 overall and improves over its initial model by +2.9, with gains on 25 of 30 benchmarks and all six categories. Among 8B models, **Vero-Qwen3I-8B** reaches 66.1 overall and outperforms Qwen3-VL-8B-Thinking by +3.8 overall and on 25 of 30 benchmarks, while **Vero-Qwen3T-8B** improves over Qwen3-VL-8B-Thinking by +3.5 overall and on 22 of 30 benchmarks. Compared with distillation warm start baselines, **Vero-Qwen25-7B** exceeds OpenMMReasoner-7B by +5.6 overall despite using a 874K example teacher warm start. We release all data, code, and models to facilitate future research.

2 Related Work

Vision-language models. Vision-language models excel on multimodal tasks, including proprietary systems such as GPT-5 [77] and Gemini [13, 81, 82], open-weight families such as Qwen [3, 4], GLM [28], and Kimi [37], and fully open releases of data, code, and weights such as Molmo [12, 15] and LLaVA [2, 46]. These models are expected to handle a wide range of tasks. While little is publicly known about proprietary model post-training, recent open-weight models have explored techniques such as RL with curriculum sampling [28] and mixed on-policy RL [106], yet the factors that drive their performance across diverse tasks

remain unclear. Our work targets this gap by providing a fully open multi-domain RL recipe for general visual understanding.

Reasoning and thinking for VLMs. Chain-of-thought reasoning enables models to use additional test-time compute through step-by-step problem decomposition [91,112]. The two dominant approaches for training reasoning models are distillation, where a strong teacher generates reasoning traces for supervised fine-tuning [69, 96, 100], and reinforcement learning, which optimizes against outcome-based rewards without a fixed teacher [14]. Recent works apply RL to visual reasoning [19,87,102,110], but primarily in narrow domains, leaving the effect of RL-trained reasoning on broad visual understanding underexplored. We show that RL with careful reward and data design consistently outperforms narrowly trained baselines across diverse visual task categories.

RL recipes and training data design for VLMs. Several works provide recipes for RL-based visual reasoning training. OpenMMReasoner [110] combines teacher distillation and GSPO [113] over multimodal reasoning benchmarks, OneThinker [19] uses a distillation warm start before RL, VL-Rethinker [87] addresses training instability via selective sample replay and forced rethinking, and Perception-R1 [102] designs discriminative rewards for perceptual tasks. These efforts target visual math or narrow perceptual domains and provide limited ablations of dataset selection, sample filtering, and reward design. Our recipe centers on **Vero-600K**, which spans six task categories with 600K data points from 59 datasets, includes a routed reward system, and provides systematic ablations of design choices, all released publicly to support open VLM research.

3 Vero

3.1 Task Categories

We consider the problem of training a Vision-Language Model (VLM) π_θ via reinforcement learning to maximize expected reward across a diverse set of visual reasoning tasks. Given a visual input v (an image or set of images) and a text query q, the model generates a structured response $y = (z, a) \sim \pi_\theta(\cdot \mid v, q)$, where z denotes the reasoning or thinking content and a denotes the final answer. We verify the answer a against the ground-truth y^*. The RL training objective is:

$$\max_\theta \; \mathbb{E}_{(v,q,y^*)\sim\mathcal{D}} \; \mathbb{E}_{(z,a)\sim\pi_\theta(\cdot|v,q)} \left[R(a, y^*) \right], \tag{1}$$

where $\mathcal{D}$ is the training data distribution. A central challenge is constructing $\mathcal{D}$ to span a broad range of visual reasoning capabilities, so that the resulting policy generalizes across diverse tasks.

Figure 2 provides an overview of our training data composition. We organize our training data into six task categories, each targeting a distinct visual reasoning capability. This taxonomy is motivated by two observations. First, we find empirically (Section 5 and Section 6) that training on any single category fails to transfer reliably to others and elicits distinct chain-of-thought behaviors, suggesting that these categories exercise different reasoning strategies and skills.

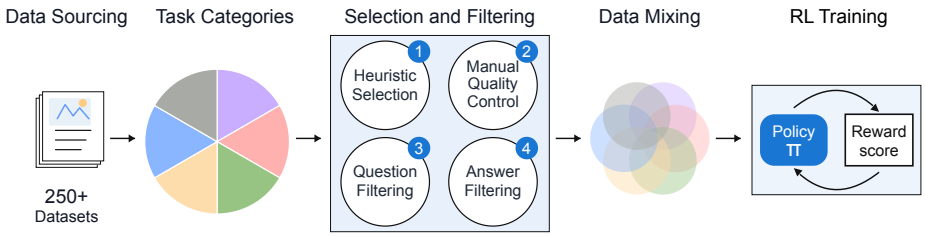

Fig. 3: Vero-600K data curation pipeline. Starting from over 250 candidate datasets, we assign each to one of six task categories and apply multi-stage selection and filtering: heuristic screening (size, resolution, answer format), manual quality control, LLM-based question filtering for ambiguity and verifiability, and answer filtering for stable reward computation. The retained data are combined into a uniformly weighted mixture across categories and used for on-policy RL with task-routed rewards.

Second, while existing VLM evaluation frameworks organize benchmarks along similar axes, e.g., Qwen2.5-VL [4] separates document understanding, mathematical reasoning, and grounding, and Kimi K2.5 [38] distinguishes reasoning from perception, these categorizations are adopted by convention rather than validated empirically, and they are designed for evaluation rather than training. Our taxonomy refines and extends these axes to cover a broader set of visual reasoning tasks, and we validate its effectiveness for multi-task RL (Section 5).

Concretely, we define six categories: **STEM** (13 datasets) covers mathematical diagram reasoning, scientific figure interpretation, and medical image understanding, with answers that are typically numeric or symbolic. **Spatial & Action** (8 datasets) targets embodied reasoning, UI navigation, and 3D spatial understanding, requiring reasoning about spatial transformations and action sequences. **Knowledge & Recognition** (12 datasets) spans visual question answering that combines object, scene, and entity recognition with external or commonsense knowledge. **Chart & OCR** (9 datasets) focuses on extracting and reasoning over structured information in documents, charts, tables, and infographics. **Grounding, Counting & Search** (11 datasets) requires spatially localizing objects via bounding boxes, counting entity instances, and searching among visual distractors. **Captioning & Instruction Following** (6 datasets) encompasses open-ended image description and following prompt instructions.

3.2 Vero-600K: Sourcing, Filtering, and Mixtures of Broad Tasks

We construct **Vero-600K**, a multi-task RL training set of 600K samples from 59 datasets spanning six task categories (Section 3.1). Figure 3 summarizes the **Vero-600K** data curation pipeline, and Figure 4 shows representative examples.

Step 1. Dataset sourcing and selection. We start from over 250 candidate datasets drawn from instruction-tuning and RL collections (e.g., FineVision [92]) and recently released task-specific sources (e.g., Visual Jigsaw [93]), then apply dataset-level filtering. Each dataset is assigned to the task category that best reflects its primary skill, based on manual inspection and its utility in prior work.

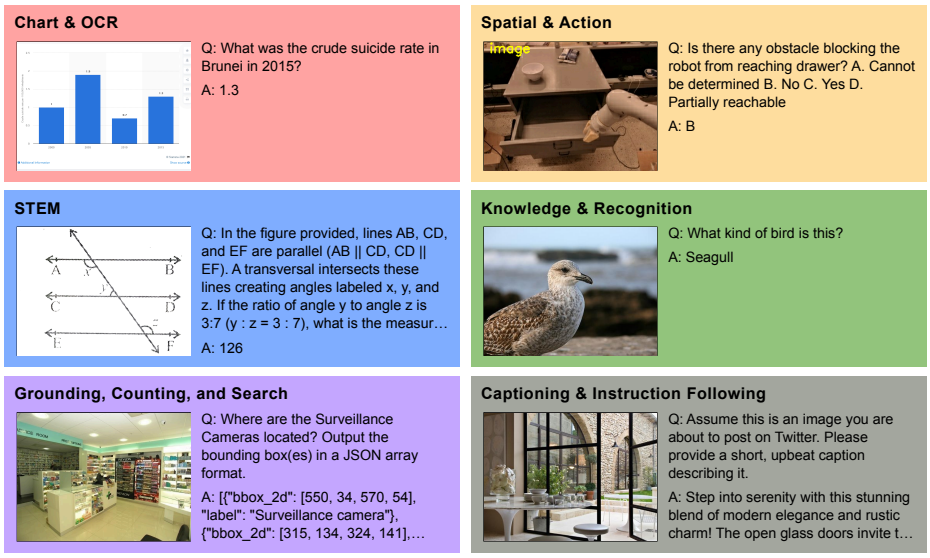

Fig. 4: Examples from each task category illustrate the breadth of Vero-600K. We show representative samples of our training data from the six categories, highlighting the diversity of visual inputs, question formats, and answer types covered by our training set.

Heuristic selection. We discard datasets with fewer than 1K examples, average image resolution below 200K pixels (retaining five low-resolution datasets for question quality), or binary questions to mitigate guessing.

Manual selection. For each candidate, we inspect ∼50 examples against three criteria: **correctness** (<5% annotation error rate in image–question–answer triples), **unambiguity** (each question admits a single verifiable answer), and **verifiability** (the answer format is compatible with our reward functions). Of ∼100 datasets passing heuristic screening, 59 were retained. For a small number of datasets, we additionally rewrite questions to fix prompt clarity (e.g., GameQA, Magma) or drop high-error subsets.

Figure E1 (Appendix) compares dataset selection strategies and shows our filtering has significant gains compared to taking a random subset from the candidate pool or sampling from the STEM or Chart subsets of the FineVision [92] training set. On STEM, our mixture achieves an average benchmark gain of +4.6, compared to +2.9 for FineVision and −0.2 for random sampling from all candidates. On Chart & OCR, the gains are +3.4, +1.9, and +2.5, respectively.

Step 2. Data filtering. After dataset-level filtering, many individual examples remain ambiguous, unanswerable, or incompatible with our rewards. We apply additional steps to filter individual prompts.

Question filtering. We use Qwen3-VL-235B-A22B-Instruct [3] to remove ambiguous, image-irrelevant, or unverifiable questions. The model scores each datapoint on five criteria: (1) *relevance*, whether the image depicts what the question

	Chart & OCR	STEM	Spatial & Action	Knowl. & Recog.	Grnd., Cnt. & Search
Unfiltered	60.0	45.4	56.3	62.5	**54.7**
Q. Filtering	**60.1**	43.6	**58.2**	63.0	54.1
A. Canonic.	59.9	**45.5**	–	**64.6**	–

Table 1: Filtering generally helps remove ambiguous samples from noisy datasets. Effect of question and answer filtering on Qwen2.5-VL-7B-Instruct (category average scores on **VeroEval**).

	Chart & OCR	STEM	Spatial & Action	Knowl. & Recog.	Grnd., Cnt. & Search	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 2: Equal task ratios perform best overall. We compare uniform, inverse-accuracy, resolution-based, reasoning-length-based, and no-Knowledge/Recognition task weighting. Values are absolute score changes (Δ) on **VeroEval** after RL from Qwen3-VL-8B-Instruct.

refers to; (2) *ambiguity*, whether the question is too vague or not a genuine question; (3) *language*, whether the question is in English; (4) *verifiability*, whether a single objectively correct answer can be derived from visible content; and (5) *numeric precision*, whether the required precision is visually unambiguous. Any triggered criterion removes the datapoint. We provide details in Appendix A.3.

Answer filtering. We normalize ground-truth answers using text-only Qwen3-235B-A22B-Instruct [3] to ensure stable reward computation. A full list of answer filtering rules is in Appendix A.4. Effects of filtering are mixed across categories (Table 1): question filtering yields a clear gain on Spatial & Action (+1.9 pts) but slightly hurts Grounding, Counting & Search (−0.6 pts). Despite this variance, we apply both steps to all applicable task categories, as they generally help remove ambiguous samples from noisy datasets.

Step 3. Data mixtures. In our multi-task RL setting, the task category sampling distribution governs how training signal is allocated across skills. We investigate four task category weighting schemes (uniform, difficulty-weighted, image-size-weighted, reasoning-length-weighted), where the number of samples per batch is determined by a ratio proportional to the metric (e.g., difficulty). Uniform sampling achieves the highest benchmark average gain (+5.8 pts over the base model), outperforming alternative schemes. Alternatives yield gains on individual categories but at the cost of others (Table 2). We use uniform task category weighting, as it achieves the best overall performance.

VeroEval evaluation suite. We introduce VeroEval, a challenging evaluation suite for broad visual reasoning. It includes 30 benchmarks across the six categories in Section 3, with three to eight benchmarks per category. Benchmarks are selected for difficulty, annotation quality, and intra-category diversity, while retaining established benchmarks such as ChartQA and ScreenSpot for comparability. The full list appears in Appendix Table A2.

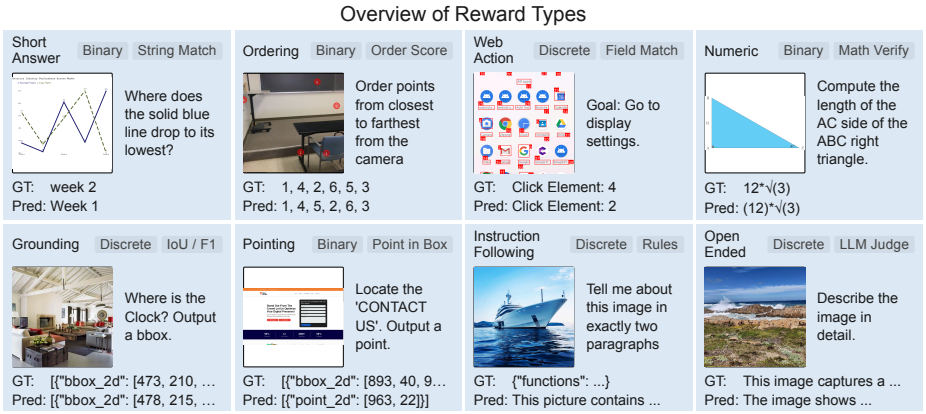

Fig. 5: Accuracy verifiers for our multi-task reward. Each card illustrates a verifier used in **Vero**, with an example visual question, ground-truth answer, and model prediction. Verifiers include binary rewards (string match, multiple choice (not displayed), list string match (not displayed), numeric, ordering, point-in-box) and graded rewards (IoU/F1 for grounding, field match for web actions, rule-based checks for instruction following, and LLM-as-judge for open-ended responses). This task-routed design enables accurate reward computation across diverse answer formats.

3.3 Training Vero with Reinforcement Learning

Algorithmic details. At its core, RL maximizes the expected reward of the model's response y given a visual input v and query q. Our RL algorithm builds on Group Relative Policy Optimization (GRPO) [74] and integrates algorithmic advances from GSPO [113], among others [103].

GSPO [113] replaces the independent per-token importance ratios of GRPO with a *sequence-level* ratio. For each response y_i in a group of G rollouts, the sequence-average log-probability difference $\bar{\Delta}_i = \frac{1}{|y_i|}\sum_t(\log \pi_\theta(y_{i,t}) - \log \pi_{\theta_{\text{old}}}(y_{i,t}))$ is used to form a token-level ratio $s_{i,t}(\theta) = \exp(\text{sg}(\bar{\Delta}_i) + \log \pi_\theta(y_{i,t}) - \text{sg}(\log \pi_\theta(y_{i,t})))$, where sg denotes stop-gradient. The GSPO objective is then:

$$\mathcal{J}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\Big(s_{i,t}(\theta)\,A_i,\ \text{clip}\big(s_{i,t}(\theta),\,1-\varepsilon_{\text{low}},\,1+\varepsilon_{\text{high}}\big)\,A_i\Big), \quad (2)$$

where $A_i = (r_i - \mu_g)/(\sigma_g + \epsilon)$ is the normalized group advantage; note that in our setting $A_{i,t} = A_i$ for all t, i.e., the advantage is constant across tokens within a response. We adopt asymmetric clip-higher [103] ($\varepsilon_{\text{high}} > \varepsilon_{\text{low}}$), remove the KL penalty [49, 103] to allow less-restricted updates, and apply a soft overlong penalty [103] that linearly ramps before the context limit.

Reward formulation. The total reward for a response y is ($\alpha = 0.2$):
$R(y, y^*) = (1-\alpha)\,R_{\text{acc}}(y, y^*) + \alpha\,R_{\text{fmt}}(y) + R_{\text{overlong}}(y)$.

Overlong penalty. To discourage excessively long responses, we use the soft penalty from [103] as a linear ramp in the buffer zone $[L_{\max} - B,\ L_{\max}]$ ($B = 2048$, $L_{\max} = \texttt{max\_tokens}$, and $\lambda = 1.0$): $R_{\text{overlong}}(y) = \min\big(-\frac{|y| - (L_{\max} - B)}{B}\,\lambda,\ 0\big)$.

Format reward. R_{fmt} requires the response to follow the format `<think>...</think> <answer>...</answer>` with non-empty think content; responses that violate this structure receive $R_{\text{fmt}} = 0$. Given valid structure, $R_{\text{fmt}} = 1$ by default. For discrete symbolic answer types (string match, multiple choice, numeric, list match, counting, ordering, search, web action), a single valid `\boxed{...}` in the answer block is additionally required for $R_{\text{fmt}} = 1$; its absence or the presence of multiple `\boxed{...}` expressions reduces R_{fmt} to 0.5.

Multi-task reward. For $R_{\text{acc}}(y, y^*)$, we detail below the ten reward functions corresponding to the task types in our dataset (Figure 5). We show in Section 4.2 that our reward design outperforms simple alternatives.

The binary rewards ($\in \{0, 1\}$) cover **string match** (normalized exact-string equality), **multiple choice** (extract a letter A–Z and compare to the prediction), **numeric** (symbolic parsing via MATH-VERIFY [39] with optional tolerance), and **list string match** (any-match over a string set, handling synonyms). The graded rewards ($\in [0, 1]$) cover **ordering** (full reward for exact order, 0.2-discounted for correct set in wrong order; Visual Jigsaw [93]), **web action** (weighted match over JSON fields `ACTION`/`MARK`/`VALUE`, scoring the fraction of non-null gold fields correct; ViGoRL [69]), **grounding** (Hungarian matching of predicted and gold boxes, IoU/F1 at threshold 0.5, coordinates normalized to $[0, 1000]$ Qwen-style; Perception-R1 [102]), **clicking** (whether the predicted point falls in the gold box; ViGoRL [69]), **instruction following** (fraction of output constraints satisfied, e.g. length, format, keywords, via the checks from MMIFeval [17] and RLVR-IFeval [62]), and **LLM-as-judge** (the OLMo3 [59] setup. The prompt scores 1–10 and penalizes meta-commentary; see Appendix D and Appendix B.3).

4 Experiments

Evaluation settings. We use model-family-specific decoding settings, following the recommended setups for Qwen2.5-VL/MiMo-VL [4] and Qwen3-VL [3]. Appendix Tables C2 and C3 list the sampling and shared runtime settings. When needed, we use Qwen3-32B with thinking disabled as the LLM judge. All models are evaluated with lmms-eval [109] using each benchmark's official protocol; benchmark-specific details appear in Appendix C. Table 3 reports Avg@5 for **Vero** variants, with superscript $\pm$ denoting standard error across runs.

Baselines. We compare against: (1) base VLMs without native <think> tokens (Qwen2.5-VL-7B-Instruct [4], Qwen3-VL-8B-Instruct [3], Qwen3.5-9B, Molmo2-O-7B [12]), (2) models trained to do native CoT with <think> tokens (Qwen3-VL-8B-Thinking [3], MiMo-VL-7B-RL [106]), (3) existing fully open RL-trained models and recipes (VL-Rethinker-7B [87], LLaVA-OV-1.5-RL [2], OpenMMReasoner-7B [110], OneThinker-8B [19]), and (4) a proprietary reasoning model gpt-5-nano-2025-08-07 [77] with medium reasoning effort. For baseline results, we prioritize scores from official technical reports and benchmark leaderboards. When published results are unavailable, we evaluate the models ourselves (indicated by †) and follow the published benchmark guidelines.

	Vero (Ours)						Open Weights Models					Fully Open RL Recipes				Propr.
	Vero Qw35-9B	Vero Qw35-9B Base	Vero Qw3I-8B	Vero Qw3T-8B	Vero Qw25-7B	Vero Mi-7B	Qw3VL 8B-Ins	Qw3VL 8B-Thk	Qw35 9B	Qw25VL 7B-Ins	MiVL RL-7B	LLaVA OV1.5 RL-8B	Open MM Reas-7B	One Thinker 8B	Mo2-O 7B	GPT-5 Nano
RL Initial Model	Qw35 9B	Qw35 9B Base	Qw3VL 8B Inst	Qw3VL 8B Think	Qw25VL 7B Inst	MiVL 7B SFT	N/A	N/A	N/A	N/A	MiVL 7B SFT	LLaVA OV1.5 Ins	Qw25vl 7B + Distill 874k	Qw3vl 8B + Distill 340k	SFT Only	N/A
Chart & OCR																
ChartQA-Pro	**71.2**$^{\pm0.1}_{+5.1}$	70.4$^{\pm0.2}_{+5.3}$	61.4$^{\pm0.2}_{+17.1}$	62.3$^{\pm0.2}_{+3.4}$	52.0$^{\pm0.2}_{+8.7}$	61.6$^{\pm0.2}_{+10.5}$	44.3†	58.9†	66.1†	43.3†	61.1†	-	48.3†	56.0†	31.4†	56.0†
ChartQA	91.7$^{\pm0.0}_{+0.1}$	92.1$^{\pm0.1}_{+4.0}$	91.7$^{\pm0.1}_{+2.1}$	90.7$^{\pm0.1}_{+2.1}$	90.3$^{\pm0.1}_{+0.0}$	90.6$^{\pm0.1}_{+4.0}$	89.6	88.6	91.3†	87.3	**94.4**	87.4	89.8†	88.0†	75.2†	80.1†
InfoVQA	90.9$^{\pm0.0}_{-0.4}$	**91.8**$^{\pm0.1}_{+14.2}$	87.5$^{\pm0.1}_{+4.4}$	88.3$^{\pm0.2}_{+2.3}$	81.6$^{\pm0.1}_{+1.0}$	87.3$^{\pm0.1}_{+5.4}$	83.1	86.0	91.3†	82.6	90.1	76.6	82.8†	74.9†	60.3†	67.9†
CharXiv$_{Reason}$	**77.4**$^{\pm0.2}_{+4.4}$	72.3$^{\pm0.2}_{+10.5}$	54.3$^{\pm0.6}_{+7.9}$	60.3$^{\pm0.3}_{+7.3}$	45.8$^{\pm0.3}_{+3.3}$	64.3$^{\pm0.5}_{+11.2}$	46.4	53.0	73.0	42.5	60.9†	-	46.1	47.5†	35.6†	51.2†
ChartMuseum	**70.3**$^{\pm0.2}_{+3.3}$	67.4$^{\pm0.3}_{+12.0}$	47.5$^{\pm0.2}_{+7.2}$	51.6$^{\pm0.3}_{+7.2}$	33.0$^{\pm0.5}_{+6.2}$	47.7$^{\pm0.4}_{+7.5}$	40.0	44.4	66.8†	26.8	48.7†	-	35.2†	35.2†	30.3†	48.0†
EvoChart	81.3$^{\pm0.1}_{+1.1}$	**81.4**$^{\pm0.1}_{+7.7}$	74.7$^{\pm0.3}_{+10.7}$	75.7$^{\pm0.1}_{+2.7}$	64.6$^{\pm0.2}_{+1.8}$	73.4$^{\pm0.2}_{+2.4}$	64.0†	73.0†	80.2†	62.8†	73.4†	-	64.4†	67.4†	51.0†	63.3†
Category Avg	**80.5**$^{\pm0.0}_{+2.3}$	79.2$^{\pm0.1}_{+8.9}$	69.5$^{\pm0.2}_{+8.3}$	71.5$^{\pm0.1}_{+1.2}$	61.2$^{\pm0.1}_{+3.6}$	70.8$^{\pm0.0}_{+6.3}$	61.2	67.3	78.1	57.6	71.4	-	61.1	61.5	47.3	61.1
STEM																
MMMU-Pro$_{Std}$	**71.3**$^{\pm0.2}_{+7.2}$	69.9$^{\pm0.3}_{+7.1}$	59.0$^{\pm0.3}_{+9.1}$	59.5$^{\pm0.2}_{+7.6}$	42.8$^{\pm0.3}_{+4.7}$	58.7$^{\pm0.2}_{+12.6}$	55.9	60.4	70.1	38.3	59.4†	39.9	44.1	53.6†	31.9†	61.3†
MMMU-Pro$_{Vis}$	**69.6**$^{\pm0.2}_{+10.9}$	67.4$^{\pm0.2}_{+12.9}$	56.6$^{\pm0.1}_{+14.5}$	57.9$^{\pm0.2}_{+4.5}$	40.0$^{\pm0.2}_{+7.7}$	53.1$^{\pm0.2}_{+5.4}$	42.1†	53.4†	58.7†	32.3†	49.8†	35.7	40.6	48.3†	16.0†	53.1†
MathVision	**79.0**$^{\pm0.2}_{+10.1}$	71.1$^{\pm0.2}_{+16.7}$	59.4$^{\pm0.2}_{+5.5}$	63.2$^{\pm0.1}_{+0.5}$	28.4$^{\pm0.1}_{+2.3}$	59.4$^{\pm0.2}_{+2.0}$	53.9	62.7	78.9†	25.1	58.8†	34.4	43.6	48.6†	21.3†	61.7†
MathVista$_{testmini}$	**86.8**$^{\pm0.4}_{+1.1}$	85.9$^{\pm0.2}_{+19.7}$	81.0$^{\pm0.2}_{+3.8}$	80.7$^{\pm0.3}_{+0.7}$	74.3$^{\pm0.3}_{+6.1}$	79.7$^{\pm0.1}_{-0.3}$	77.2	81.4	85.7	68.2	80.4†	72.3	79.5	77.6	53.6†	70.2†
Category Avg	**76.7**$^{\pm0.1}_{+3.5}$	73.6$^{\pm0.2}_{+14.1}$	64.0$^{\pm0.0}_{+6.7}$	65.3$^{\pm0.1}_{+0.8}$	46.4$^{\pm0.2}_{+5.4}$	62.7$^{\pm0.1}_{+2.4}$	57.3	64.5	73.3	41.0	62.1	45.6	52.0	57.0	30.7	61.6
Spatial & Action																
Blink	**70.9**$^{\pm0.2}_{+2.5}$	69.7$^{\pm0.4}_{+4.6}$	68.3$^{\pm0.2}_{+0.6}$	66.6$^{\pm0.3}_{+1.9}$	58.6$^{\pm0.1}_{+2.2}$	63.3$^{\pm0.1}_{+0.0}$	69.1	64.7	67.4†	56.4	64.5†	-	60.8†	61.9†	56.4†	59.3†
ERQA	**55.8**$^{\pm0.2}_{+0.2}$	52.5$^{\pm0.7}_{+4.5}$	47.6$^{\pm0.4}_{+1.9}$	46.2$^{\pm0.5}_{+0.5}$	42.4$^{\pm0.6}_{+0.6}$	40.9$^{\pm0.4}_{+1.4}$	45.8	46.8	55.5	41.8†	43.5†	-	39.2†	43.2†	43.5†	45.5†
GameQA$_{Lite}$	**75.3**$^{\pm0.2}_{+13.5}$	68.3$^{\pm0.2}_{+18.5}$	52.5$^{\pm0.3}_{+18.5}$	54.9$^{\pm0.3}_{+15.1}$	45.3$^{\pm0.1}_{+19.2}$	52.0$^{\pm0.1}_{+2.8}$	34.0†	39.8†	61.4†	26.1†	49.8†	-	29.2†	41.4†	29.6†	45.9†
EmbSpatial	**84.0**$^{\pm0.2}_{+1.0}$	81.8$^{\pm0.2}_{+3.6}$	79.4$^{\pm0.2}_{+0.9}$	80.8$^{\pm0.1}_{+0.3}$	69.8$^{\pm0.1}_{+0.9}$	71.4$^{\pm0.1}_{-1.1}$	78.5	81.1	83.0	70.7†	70.2†	-	71.0†	72.0†	68.1†	74.2†
CV Bench	89.0$^{\pm0.1}_{+1.2}$	**89.9**$^{\pm0.2}_{+2.6}$	88.1$^{\pm0.1}_{+2.6}$	87.9$^{\pm0.1}_{+1.9}$	81.4$^{\pm0.1}_{+1.0}$	83.7$^{\pm0.1}_{+1.1}$	85.5†	86.0†	87.5†	80.4†	83.5†	82.9	81.1†	82.9†	81.7†	82.5†
Category Avg	**75.0**$^{\pm0.0}_{+4.0}$	72.4$^{\pm0.3}_{+6.8}$	67.2$^{\pm0.1}_{+4.6}$	67.3$^{\pm0.2}_{+3.6}$	59.5$^{\pm0.1}_{+1.4}$	62.3$^{\pm0.1}_{+1.6}$	62.6	63.7	71.0	55.1	62.3	-	56.3	60.3	55.9	61.5
Knowledge & Recognition																
RealWorldQA	80.1$^{\pm0.0}_{+0.2}$	78.3$^{\pm0.4}_{+6.5}$	74.2$^{\pm0.4}_{+2.7}$	71.4$^{\pm0.2}_{+0.1}$	68.7$^{\pm0.1}_{+0.2}$	70.0$^{\pm0.4}_{+2.6}$	71.5	73.5	**80.3**	68.5	68.6†	68.4	69.9†	68.2†	73.3†	65.9†
SimpleVQA$_{En}$	49.5$^{\pm0.5}_{+4.4}$	**53.9**$^{\pm0.3}_{+12.2}$	45.9$^{\pm0.3}_{+1.7}$	43.9$^{\pm0.3}_{+1.9}$	50.1$^{\pm0.3}_{+5.6}$	46.4$^{\pm0.2}_{+4.5}$	44.2†	44.9†	44.6†	44.5†	40.9†	-	43.8†	43.1†	30.8†	36.6†
FVQA	**32.4**$^{\pm1.3}_{+1.3}$	26.5$^{\pm1.2}_{+3.2}$	24.3$^{\pm0.4}_{+3.7}$	22.1$^{\pm0.2}_{+2.1}$	26.6$^{\pm0.1}_{+0.7}$	27.9$^{\pm0.3}_{+1.9}$	26.0	24.2	31.4†	21.9	31.8†	-	21.4†	22.8†	17.7†	29.4†
MM-Vet v2	81.0$^{\pm0.8}_{+4.3}$	**84.1**$^{\pm0.3}_{+11.5}$	71.0$^{\pm0.3}_{+3.4}$	82.4$^{\pm0.4}_{+7.9}$	66.1$^{\pm0.4}_{+2.2}$	78.5$^{\pm0.4}_{+18.8}$	67.6	74.5	76.5†	62.9	61.2†	-	62.6†	66.3†	60.8†	71.5†
Category Avg	**60.7**$^{\pm0.0}_{+2.5}$	60.7$^{\pm0.1}_{+7.2}$	53.8$^{\pm0.2}_{+1.5}$	54.9$^{\pm0.2}_{+0.6}$	52.8$^{\pm0.1}_{+3.3}$	55.7$^{\pm0.1}_{+5.9}$	52.3	54.3	58.2	49.5	50.6	-	49.4	50.1	45.6	50.9
Grounding, Counting & Search																
CountBenchQA	95.5$^{\pm0.2}_{+0.0}$	94.1$^{\pm0.2}_{+3.7}$	90.0$^{\pm0.4}_{+1.2}$	92.1$^{\pm0.5}_{+2.5}$	83.1$^{\pm0.4}_{-2.8}$	84.2$^{\pm0.3}_{-1.1}$	88.8†	89.6†	**95.5**†	85.9†	86.4†	86.8	83.7†	89.2†	89.4†	75.4†
CountQA	60.7$^{\pm0.2}_{+1.1}$	48.6$^{\pm0.4}_{+12.4}$	34.6$^{\pm0.2}_{+6.1}$	37.0$^{\pm0.4}_{+2.2}$	24.1$^{\pm0.2}_{+2.2}$	26.5$^{\pm0.1}_{+0.9}$	28.5†	31.8†	**61.8**†	20.9†	27.4†	-	22.0†	28.8†	32.1†	25.7†
MME-RealWorld-Lite	61.7$^{\pm0.2}_{+0.7}$	**61.8**$^{\pm0.1}_{+5.1}$	57.5$^{\pm0.2}_{+10.4}$	54.4$^{\pm0.1}_{+0.6}$	53.6$^{\pm0.1}_{+0.1}$	52.8$^{\pm0.2}_{+9.0}$	47.1†	44.8†	61.0†	44.5†	48.7†	-	53.9†	42.6†	44.4†	49.8†
VStarBench	**90.5**$^{\pm0.3}_{+0.4}$	88.6$^{\pm0.3}_{+3.8}$	88.7$^{\pm0.5}_{+6.5}$	82.9$^{\pm0.6}_{+0.5}$	84.8$^{\pm0.0}_{+5.7}$	83.8$^{\pm0.0}_{+1.6}$	82.2†	76.4†	90.1	80.1†	84.3	79.1	81.7†	66.5†	73.8†	71.2†
AerialVG	**45.3**$^{\pm0.1}_{+11.2}$	42.3$^{\pm0.0}_{+20.1}$	30.2$^{\pm0.1}_{+2.4}$	32.3$^{\pm0.2}_{+10.7}$	29.1$^{\pm0.0}_{+6.8}$	15.5$^{\pm0.1}_{-6.6}$	32.2†	12.6†	34.1†	22.3†	22.2†	-	11.9†	24.8†	-	-
VisualProbe	55.0$^{\pm0.4}_{+2.8}$	**55.9**$^{\pm0.2}_{+0.2}$	52.0$^{\pm0.5}_{+4.3}$	46.3$^{\pm0.4}_{+7.0}$	52.1$^{\pm0.0}_{+6.1}$	51.6$^{\pm0.0}_{+4.8}$	47.7†	39.3†	52.2†	46.0†	52.2†	-	45.9†	21.3†	34.8†	41.5†
ScreenSpot	90.8$^{\pm0.3}_{+5.3}$	**93.9**$^{\pm0.1}_{+61.0}$	92.6$^{\pm0.1}_{+6.0}$	91.7$^{\pm0.2}_{+0.2}$	90.2$^{\pm0.0}_{+4.6}$	90.6$^{\pm0.1}_{+2.7}$	86.6†	85.5†	85.5†	85.6†	87.3†	-	63.5†	84.2†	75.8†	-
ScreenSpotPro	64.2$^{\pm0.2}_{+7.3}$	**68.0**$^{\pm0.1}_{+59.3}$	61.8$^{\pm0.2}_{+7.2}$	48.7$^{\pm0.4}_{+2.1}$	39.6$^{\pm0.0}_{+10.7}$	36.9$^{\pm0.0}_{+4.1}$	54.6	46.6	65.2	23.9†	37.4†	-	5.8†	20.7†	19.3†	-
Category Avg	**70.5**$^{\pm0.1}_{+2.5}$	69.1$^{\pm0.1}_{+23.0}$	63.4$^{\pm0.1}_{+4.9}$	60.7$^{\pm0.1}_{+7.4}$	57.1$^{\pm0.0}_{+6.0}$	55.2$^{\pm0.0}_{+2.2}$	58.5	53.3	68.2	51.1	55.7	-	46.0	47.3	-	-
Captioning & IF																
MM-MTBench	83.8$^{\pm0.4}_{+0.7}$	**85.8**$^{\pm0.4}_{+14.3}$	81.0$^{\pm0.3}_{+6.6}$	71.4$^{\pm0.6}_{-0.9}$	62.6$^{\pm0.5}_{+0.7}$	75.1$^{\pm0.7}_{-0.6}$	74.4	77.8	78.1†	58.9	79.2†	-	37.2†	60.4†	33.3†	72.7†
MIABench	92.0$^{\pm0.1}_{+0.2}$	**93.5**$^{\pm0.2}_{+2.5}$	93.4$^{\pm0.1}_{+2.3}$	92.4$^{\pm0.3}_{+0.3}$	87.2$^{\pm0.1}_{+2.4}$	90.7$^{\pm0.2}_{+2.3}$	91.1	91.5	89.8†	81.8	88.4†	-	66.6†	81.4†	77.9†	92.0†
MMIFEval	**86.4**$^{\pm0.3}_{+4.3}$	84.0$^{\pm0.3}_{+11.3}$	76.0$^{\pm0.5}_{+6.8}$	78.3$^{\pm0.3}_{+1.3}$	66.6$^{\pm0.4}_{+13.0}$	78.7$^{\pm0.4}_{+9.5}$	69.2	75.0	82.1†	53.6	66.1	-	39.7†	53.5†	54.4†	78.0†
Category Avg	87.4$^{\pm0.1}_{+1.4}$	**87.8**$^{\pm0.1}_{+9.8}$	83.5$^{\pm0.2}_{+5.3}$	80.7$^{\pm0.3}_{+0.7}$	72.1$^{\pm0.3}_{+7.3}$	81.5$^{\pm0.2}_{+3.7}$	78.2	81.4	83.3	64.8	77.9	-	47.8	65.1	55.2	80.9
Overall Avg	**74.4**$^{\pm0.0}_{+2.9}$	73.0$^{\pm0.0}_{+12.9}$	66.1$^{\pm0.0}_{+5.4}$	65.8$^{\pm0.1}_{+3.5}$	57.8$^{\pm0.1}_{+4.9}$	63.2$^{\pm0.1}_{+3.6}$	60.7	62.3	71.5	52.9	62.4	-	52.2	55.7	-	-

Table 3: Vero achieves state-of-the-art performance across six task categories on VeroEval. Vero columns show Avg@5 over initial models trained with RL on our dataset; superscript ± values report the standard error of the mean across runs, and +x / -x deltas indicate improvement/decline over the respective initial model. In model names, Qw denotes Qwen and Mi denotes MiMo. † indicates results evaluated by us. All other results are taken from official technical reports.

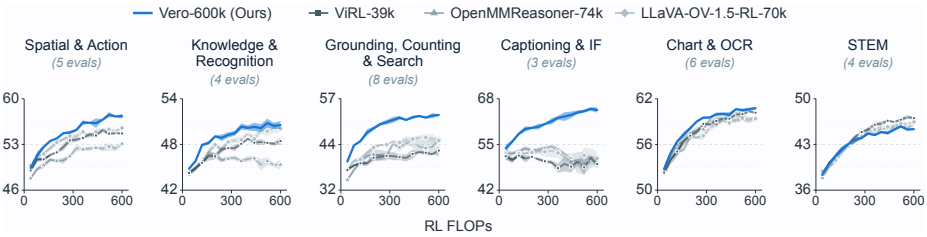

Fig. 6: Per-category RL training curves. Evaluation score vs. RL FLOPs for RL runs using **Vero-600K** and three prior open RL datasets, all starting from Qwen2.5-VL-7B-Instruct. Curves report mean and SEM across seeds where shown. Dashed segments indicate training beyond one epoch. **Vero-600K** leads on five of six categories throughout training; on STEM it remains within 2 points of the best prior dataset.

4.1 Evaluation Results on VeroEval

We report results in Table 3 and highlight the following observations.

Strong performance with a fully open RL recipe. **Vero-Qwen35-9B** achieves the highest overall average in Table 3, reaching 74.4 on **VeroEval**. It improves over Qwen3.5-9B by +2.9 overall, wins 25 of 30 benchmarks, and improves all six category averages: Chart & OCR (+2.3), STEM (+3.3), Spatial & Action (+4.0), Knowledge & Recognition (+2.5), Grounding, Counting & Search (+2.3), and Captioning & IF (+4.1). Among 8B models, **Vero-Qwen3I-8B** and **Vero-Qwen3T-8B** reach 66.1 and 65.8 overall, respectively, outperforming Qwen3-VL-8B-Thinking by +3.8 and +3.5 overall. Against distillation warm start baselines using the same initial model, **Vero-Qwen25-7B** is higher than OpenMMReasoner-7B by +5.6 overall, wins 5 of 6 category averages, and has its largest margins on Captioning & IF (+24.3) and Grounding, Counting & Search (+11.1). **Vero-Qwen3I-8B** is higher than OneThinker-8B [19] by +10.4 overall and wins all six category averages. **Vero** uses no additional teacher distillation.

Consistent gains across initial models. **Vero** improves all six initial models, with +2.9 to +5.4 gains across the five post-trained checkpoints. **Vero-Qwen3I-8B** improves Qwen3-VL-8B-Instruct by +5.4 overall and outperforms Qwen3-VL-8B-Thinking on 25/30 benchmarks without distillation. **Vero-Qwen3T-8B** improves Qwen3-VL-8B-Thinking by +3.5 and on 22/30 benchmarks, and **Vero-Qwen25-7B** improves Qwen2.5-VL-7B-Instruct by +4.9 and on 27/30 benchmarks. The largest gain comes from applying RL on Qwen3.5-9B-Base: +12.9 overall, reaching 73.0, the second-highest score in Table 3, without SFT or distillation. Gains concentrate on Grounding, Counting & Search (+23.0), Captioning & IF (+9.8, table's best at 87.8), and STEM (+14.1).

Vero-MiMo-7B, trained on MiMo-VL-7B-SFT with our fully open recipe, improves by +3.6 overall. It is +0.8 overall above MiMo-VL-7B-RL, which trains on the same initial model but uses a proprietary RL recipe with non-public data. **Vero-MiMo-7B** is higher on STEM (+0.6), Knowledge & Recognition (+5.1),

(a) SFT vs. RL

	Chart & OCR	STEM	Spatial & Action	Knowl. & Recog.	Grnd., Cnt. & Search	Cap. & IF	Overall Avg.
Base	57.6	41.0	55.1	49.4	50.1	64.8	52.4
FineVis SFT	54.8	37.4	52.1	45.3	40.1	52.2	46.2
Vero SFT	52.5	40.1	58.1	50.8	52.7	64.1	52.8
Vero RL	**61.9**	**46.7**	**59.4**	**53.5**	**55.0**	**70.6**	**57.2**

(b) Reward design

	Chart & OCR	STEM	Spatial & Action	Knowl. & Recog.	Grnd., Cnt. & Search	Cap. & IF	Overall Avg.
Base	57.6	41.0	55.1	49.4	50.1	64.8	52.4
Math Ver.	61.4	46.1	58.5	50.1	51.0	34.3	51.8
Ours	**61.9**	**46.7**	**59.4**	**53.5**	**55.0**	**70.6**	**57.2**

(c) RL algorithm (1/4 epoch, 5 task categories)

	Chart & OCR	STEM	Spatial & Action	Knowl. & Recog.	Grnd., Cnt. & Search	Avg.	Avg. Entropy
DAPO	58.9	45.3	57.1	49.7	52.2	54.3	$0.22_{\pm 0.15}$
GRPO	**59.2**	44.4	58.1	48.2	**53.0**	54.3	$0.50_{\pm 0.11}$
GSPO	59.0	**45.4**	**58.4**	**50.4**	**53.0**	**54.7**	$\mathbf{0.58}_{\pm 0.11}$

Table 4: Ablation studies. All results on Qwen2.5-VL-7B-Instruct. Tables (a)–(c) report absolute scores. All runs are trained 1 epoch on the 600K mixture unless specified.

and Captioning & IF (+3.6), tied on Spatial & Action, and lower on Chart & OCR and Grounding, Counting & Search.

Improvements not limited to a single domain. Unlike prior open RL-trained VLMs that focus primarily on STEM, **Vero** yields improvements across all six task categories. For example, **Vero-Qwen3I-8B** improves over its initial model on Chart & OCR (+8.3), STEM (+6.7), Spatial & Action (+4.6), Knowledge & Recognition (+1.5), Grounding, Counting & Search (+4.9), and Captioning & IF (+5.3), demonstrating that multi-task RL produces broadly capable models rather than specialists.

Consistent advantage over prior open RL datasets across training. Figure 6 compares RL runs using **Vero-600K** against runs using three prior open RL datasets, ViRL-39k [87], OpenMMReasoner-74k [110], and LLaVA-OV-1.5-RL-70k [2]. All runs start from Qwen2.5-VL-7B-Instruct and use identical RL settings over 600 steps, averaged over three runs with different random seeds. Even in the early training regime (first ∼150 steps), where all datasets are within their first epoch, **Vero-600K** already leads or remains within seed variance of the best prior dataset on every category. By the end of training, **Vero-600K** reaches the highest score on five of six categories, with the largest margins on Captioning & IF (+13.7 over the next best) and Grounding, Counting & Search (+6.7). On STEM, where prior datasets concentrate their selection, **Vero-600K** remains competitive (45.3 vs. 47.0 for ViRL-39k and 46.2 for OpenMMReasoner-74k), trailing by fewer than 2 points despite learning across categories.

4.2 Ablations

Multi-task RL requires more expressive reward design. We compare our multi-route reward design against math_verify [39], a widely used reward that

Fig. 7: Diverse task mixing eliminates negative cross-task transfer. Each row shows a model trained on a single task category (or mixture of all categories). Values are absolute score changes relative to the base model. Single-task training yields selective transfer, while mixing achieves consistent gains.

performs extraction, parsing, and grading. Results in Table 4(b) show that our reward design, which routes answers through type-specific comparisons (exact match, numeric tolerance, set matching, and LLM-judge evaluation), achieves stronger performance than math_verify across task categories. math_verify lacks the flexibility to handle the diverse answer formats.

Our data benefits most from RL, yet even with SFT alone it outperforms strong SFT baselines. We compare SFT and RL training on our dataset in Table 4(a). The SFT model is trained to directly output the final answer without chain-of-thought or `<think>` tokens. SFT on our dataset produces gains on most tasks and outperforms SFT on a recent post-training dataset, FineVision [92]. However, RL (GSPO with our multi-route reward) yields more consistent improvements across all task categories.

GSPO outperforms GRPO and DAPO and leads to more stable entropy. We compare three RL algorithms (DAPO, GRPO, and GSPO) using the same base model (Qwen2.5-VL-7B-Instruct), reward design, and training dataset. Consistent with recent findings [110], results in Table 4(c) show that GSPO achieves the highest average score (54.7) across all task categories, outperforming both GRPO (54.3) and DAPO (54.3). GSPO also maintains substantially more stable entropy throughout training (0.58 ± 0.11) compared to GRPO (0.50 ± 0.11) and especially DAPO (0.22 ± 0.15), suggesting that GSPO's sequence-level clipping better preserves exploration capacity and avoids premature policy collapse observed with alternative algorithms.

5 Data Diversity & Cross-Task Transfer

We study cross-task generalization by training models on each individual task category (100K samples, 1 epoch) and evaluating across all six categories. We compare against a model trained on a mixture of all categories with the same

total number of training samples (100K, compute-controlled) and the full dataset (600K). We report results on two base models in Figure 7.

Single-task training frequently produces neutral or negative transfer on non-target tasks. On Qwen2.5-VL, nearly all single-task-category models degrade Grounding, Counting & Search performance (e.g., −3.2 from Chart & OCR, −3.3 from STEM, −4.3 from Spatial & Action), and training on Captioning & Instruction Following alone reduces performance on all other categories (−2.1 to −7.7). Conversely, training on any non-captioning task category severely degrades Captioning & Instruction Following (−15.8 to −35.5 on Qwen2.5-VL). These patterns hold on Qwen3-VL, where single-task-category models similarly hurt Grounding (−0.8 to −5.1 for non-grounding) and Captioning & Instruction Following (−6.9 to −13.9). However, we do observe selective positive transfer: STEM improves Chart & OCR (+3.6 on Qwen3-VL), and Spatial & Action training yields gains on STEM (+3.5 on Qwen2.5-VL, +6.6 on Qwen3-VL).

Diverse task mixing eliminates negative cross-task transfer. Under the same compute budget, the mixed model achieves positive gains across all categories on both base models (+0.3 to +4.2 on Qwen2.5-VL; +1.9 to +5.2 on Qwen3-VL), avoiding the losses from single-domain training. The 600K mixture further amplifies these gains. This consistency suggests that multi-task RL is important for broad capability. Section 6 provides further evidence that reasoning behaviors may not transfer readily because category chain-of-thought patterns are distinct. We also show in Appendix D that prompting or format fixes, without open-ended training, do not prevent visual-chat degradation.

Task categories elicit markedly different reasoning lengths. Appendix Figure D1 shows average reasoning length for Qwen3-VL-8B-Instruct trained on each task category. Spatial & Action produces the longest responses (1983.3±50.8 words), followed by Chart & OCR (1592.7 ± 32.5) and STEM (1576.1 ± 39.6). Captioning & Instruction Following is much shorter (413.8±13.1), while Grounding, Counting & Search and Knowledge & Recognition are shortest (124.9 ± 12.6 and 75.8 ± 2.9). The $> 26\times$ gap between Spatial & Action and Knowledge & Recognition suggests that long chain-of-thought behavior concentrates in tasks requiring multi-step state tracking or structured decomposition.

Broader exposure to the mixed training distribution yields continued gains. Appendix Figure D3 tracks performance during a single pass over the 600K-sample mixture for three **Vero** variants (**Vero-Qwen25-7B**, **Vero-MiMo-7B**, and **Vero-Qwen3T-8B**) on CharXiv, CountBenchQA, and MMIFEval, so later points reflect greater exposure to diverse RL samples. From the 100K checkpoint to the final checkpoint, all nine model–benchmark curves improve, with a mean gain of +4.3 points. The largest late-stage gains appear on CharXiv Reasoning (mean +5.7 pp across models; up to +7.3 pp for **Vero-MiMo-7B**) and MMIFEval (mean +5.5 pp; up to +6.1 pp for **Vero-Qwen25-7B**), while CountBenchQA improves more modestly (mean +1.6 pp), having largely plateaued by the 100K checkpoint. The same late-stage trend holds on the broader evaluation suite (not shown), with continued gains on benchmarks such as ScreenSpot-Pro, GameQA Lite, and ChartMuseum and near-saturation

on MMMU-Pro Vision. These trends indicate that exposing the policy to a more diverse training distribution remains beneficial over long training.

6 Chain-of-Thought Behaviors

Benchmark accuracy alone does not characterize how the trained models arrive at their answers. We analyze the generated reasoning traces at two levels: high-level cognitive behaviors and lower-level recurring skills. Together, these analyses quantify whether models trained on different task categories exhibit consistent differences in their intermediate reasoning traces.

Experimental Setup. We evaluate model behavior using the cognitive framework of [32], which defines 28 textual reasoning behaviors, supplemented with six behaviors for visual analysis. We compute the presence rate of each behavior for every trained model from Section 5 across all six validation task categories. Behavior presence rate on **Vero** trained on Qwen3-VL-8B-Instruct is shown in Appendix Figure E3, and we report additional base models in Appendix E.2.

Each task category elicits a distinct cognitive behavioral profile. The behavioral profiles reveal that training on each task category elicits different cognitive behaviors. Captioning often uses mental imagery simulation (0.64 vs. 0.57 cross-task-category average in Qwen3), chart-trained models trigger systematic regional synthesis (0.74 vs. 0.68), and spatial reasoning often uses perception-then-reasoning sequencing (0.84 vs. 0.73). Grounding tasks more often suppress introspective behaviors, with self-awareness dropping to 0.49 (vs. 0.73), redirecting capacity toward directed visual search. In contrast, task categories requiring multi-step integration elicit higher-order behaviors, with STEM tasks showing elevated backtracking (0.48 vs. 0.27). The mixed-task-category setting increases strategy selection (0.80 vs. 0.71), indicating that the model first selects a reasoning approach before executing it.

Conclusions. We presented **Vero**, a fully open vision-language reasoning family trained with single-stage RL on **Vero-600K**, a 600K-sample dataset from 59 datasets spanning six capability categories. Our central finding is that breadth, not specialization, drives general visual reasoning: training jointly across diverse task categories with balanced, uniform exposure and task-routed rewards yields consistent positive cross-category transfer, both in benchmark performance and in the chain-of-thought behaviors models acquire. These results recast multi-task RL as a problem of distribution design, and they show that visual reasoning is not a monolithic skill but a composite of task-dependent behavioral modes that a single training mixture can elicit together. By releasing the full recipe, we aim to make this recipe a reproducible foundation for future work on open visual reasoning, and to invite further study of which task taxonomies, scales, and behaviors most effectively compose into general capability.

Acknowledgements. This work was supported by Princeton Research Computing. We thank Princeton Language and Intelligence (PLI) for their support of this research, as well as Google's TPU Research Cloud (TRC) program.

References

1. Acharya, M., Kafle, K., Kanan, C.: Tallyqa: Answering complex counting questions. In: AAAI (2019)
2. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., et al.: Llava-onevision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Berke, K.: keremberke/indoor-scene-classification via datasets at hugging face (2024)
6. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusiñol, M., Valveny, E., Jawahar, C.V., Karatzas, D.: Scene text visual question answering. In: ICCV (2019)
7. Cao, J., Xiao, J.: An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In: COLING (2022)
8. Chen, K., Xie, S., Ma, Z., Sanketi, P.R., Goldberg, K.: Robo2vlm: Improving visual question answering using large-scale robot manipulation data. In: NeurIPS (2025)
9. Cheng, K., Sun, Q., Chu, Y., Xu, F., YanTao, L., Zhang, J., Wu, Z.: Seeclick: Harnessing gui grounding for advanced visual gui agents. In: ACL (2024)
10. Cheng, X., Zhang, W., Zhang, S., Yang, J., Guan, X., Wu, X., Li, X., Zhang, G., Liu, J., Mai, Y., et al.: Simplevqa: Multimodal factuality evaluation for multimodal large language models. In: ICCV (2025)
11. Cheng, Z., Hu, J., Liu, Z., Si, C., Li, W., Gong, S.: V-star: Benchmarking video-llms on video spatio-temporal reasoning. arXiv preprint arXiv:2503.11495 (2025)
12. Clark, C., Zhang, J., Ma, Z., Park, J.S., Salehi, M., Tripathi, R., Lee, S., Ren, Z., Kim, C.D., Yang, Y., et al.: Molmo2: Open weights and data for vision-language models with video understanding and grounding. arXiv preprint arXiv:2601.10611 (2026)
13. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
14. DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., et al.: DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. Nature (2025)
15. Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., Lu, J., Anderson, T., Bransom, E., Ehsani, K., Ngo, H., Chen, Y., Patel, A., Yatskar, M., Callison-Burch, C., Head, A., Hendrix, R., Bastani, F., VanderBilt, E., Lambert, N., Chou, Y., Chheda, A., Sparks, J., Skjonsberg, S., Schmitz, M., Sarnat, A., Bischoff, B., Walsh, P., Newell, C., Wolters, P., Gupta, T., Zeng, K.H., Borchardt, J., Groeneveld, D., Nam, C., Lebrecht, S., Wittlif, C., Schoenick, C., Michel, O., Krishna, R., Weihs, L., Smith, N.A., Hajishirzi, H., Girshick, R., Farhadi, A., Kembhavi, A.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In: CVPR (2025)

16. Didolkar, A., Ballas, N., Arora, S., Goyal, A.: Metacognitive reuse: Turning recurring llm reasoning into concise behaviors. arXiv preprint arXiv:2509.13237 (2025)
17. Ding, S., Wu, S., Zhao, X., Zang, Y., Duan, H., Dong, X., Zhang, P., Cao, Y., Lin, D., Wang, J.: Mm-ifengine: Towards multimodal instruction following. In: ICCV (2025)
18. Du, M., Wu, B., Li, Z., Huang, X.J., Wei, Z.: Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In: ACL (2024)
19. Feng, K., Zhang, M., Li, H., Fan, K., Chen, S., Jiang, Y., Zheng, D., Sun, P., Zhang, Y., Sun, H., et al.: Onethinker: All-in-one reasoning model for image and video. arXiv preprint arXiv:2512.03043 (2025)
20. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: ECCV (2024)
21. Gao, J., Pi, R., Zhang, J., Ye, J., Zhong, W., Wang, Y., Hong, L., Han, J., Xu, H., Li, Z., Kong, L.: G-llava: Solving geometric problem with multi-modal large language model. In: ICLR (2025)
22. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR (2017)
23. Guha, E., Marten, R., Keh, S., Raoof, N., Smyrnis, G., Bansal, H., Nezhurina, M., Mercat, J., Vu, T., Sprague, Z., et al.: Openthoughts: Data recipes for reasoning models. In: ICLR (2026)
24. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: CVPR (2018)
25. He, W., Xi, Z., Zhao, W., Fan, X., Ding, Y., Shan, Z., Gui, T., Zhang, Q., Huang, X.: Distill visual chart reasoning ability from llms to mllms. In: Findings of EMNLP (2025)
26. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
27. Hessel, M., Soyer, H., Espeholt, L., Czarnecki, W., Schmitt, S., Van Hasselt, H.: Multi-task deep reinforcement learning with popart. In: AAAI (2019)
28. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., et al.: GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025)
29. Huang, M., Lai, H., Zhang, X., Wu, W., Ma, J., Zhang, L., Liu, J.: Evochart: A benchmark and a self-training approach towards real-world chart understanding. In: AAAI (2025)
30. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019)
31. Jia, Y., Li, J., Yue, X., Li, B., Nie, P., Zou, K., Chen, W.: Visualwebinstruct: Scaling up multimodal instruction data through web search. In: EMNLP (2025)
32. Kargupta, P., Li, S.S., Wang, H., Lee, J., Chen, S., Ahia, O., Light, D., Griffiths, T.L., Kleiman-Weiner, M., Han, J., Celikyilmaz, A., Tsvetkov, Y.: Cognitive foundations for reasoning and their manifestation in llms. arXiv preprint arXiv:2511.16660 (2025)
33. Kazemi, M., Alvari, H., Anand, A., Wu, J., Chen, X., Soricut, R.: Geomverse: A systematic evaluation of large models for geometric reasoning. arXiv preprint arXiv:2312.12241 (2023)

34. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferItGame: Referring to objects in photographs of natural scenes. In: EMNLP (2014)
35. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: ECCV (2016)
36. Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., Hajishirzi, H.: Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In: CVPR (2017)
37. Kimi Team: Kimi-VL technical report. arXiv preprint arXiv:2504.07491 (2025)
38. Kimi Team, Bai, T., Bai, Y., Bao, Y., et al.: Kimi K2.5: Visual agentic intelligence. arXiv preprint arXiv:2602.02276 (2026)
39. Kydlíček, H.: Math-verify: Math verification library. https://github.com/huggingface/Math-Verify (2025)
40. Lai, X., Li, J., Li, W., Liu, T., Li, T., Zhao, H.: Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. In: ICLR (2026)
41. Lau, J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific Data (2018)
42. Lerner, P., Ferret, O., Guinaudeau, C., Le Borgne, H., Besançon, R., Moreno, J.G., Lovon, J.: ViQuAE, a Dataset for Knowledge-based Visual Question Answering about Named Entities. In: SIGIR (2022)
43. Li, A., Wang, C., Fu, D., Yue, K., Cai, Z., Zhu, W.B., Liu, O., Guo, P., Neiswanger, W., Huang, F., Goldstein, T., Goldblum, M.: Zebra-cot: A dataset for interleaved vision-language reasoning. In: ICLR (2026)
44. Li, K., Meng, Z., Lin, H., Luo, Z., Tian, Y., Ma, J., Huang, Z., Chua, T.S.: Screenspot-pro: Gui grounding for professional high-resolution computer use. In: MM (2025)
45. Li, L., Wang, Y., Xu, R., Wang, P., Feng, X., Kong, L., Liu, Q.: Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In: ACL (2024)
46. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
47. Liu, J., Chen, Q., Wang, Z., Tang, Y., Zhang, Y., Yan, C., Wang, D., Li, X., Zhao, B.: Aerialvg: A challenging benchmark for aerial visual grounding by exploring positional relations. In: ICCV (2025)
48. Liu, Y., Zhang, B., Zang, Y., Cao, Y., Xing, L., Dong, X., Duan, H., Lin, D., Wang, J.: Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. In: CVPR (2026)
49. Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W.S., Lin, M.: Understanding r1-zero-like training: A critical perspective. In: COLM (2025)
50. Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H.W., Tay, Y., Zhou, D., Le, Q.V., Zoph, B., Wei, J., et al.: The flan collection: Designing data and methods for effective instruction tuning. In: ICML (2023)
51. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In: ICLR (2024)
52. Lu, P., Qiu, L., Chen, J., Xia, T., Zhao, Y., Zhang, W., Yu, Z., Liang, X., Zhu, S.C.: Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In: NeurIPS (2021)
53. Ma, X., Ding, Z., Luo, Z., Chen, C., Guo, Z., Wong, D.F., Feng, X., Sun, M.: Deepperception: Advancing r1-like cognitive visual perception in mllms for knowledge-intensive visual grounding. arXiv preprint arXiv:2503.12797 (2025)

54. Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of ACL (2022)
55. Masry, A., Islam, M.S., Ahmed, M., Bajaj, A., Kabir, F., Kartha, A., Laskar, M.T.R., Rahman, M., Rahman, S., Shahmohammadi, M., et al.: Chartqapro: A more diverse and challenging benchmark for chart question answering. In: Findings of ACL (2025)
56. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: WACV (2022)
57. Meng, F., Du, L., Liu, Z., Zhou, Z., Lu, Q., Fu, D., Han, T., Shi, B., Wang, W., He, J., Zhang, K., Luo, P., Qiao, Y., Zhang, Q., Shao, W.: Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. arXiv preprint arXiv:2503.07365 (2025)
58. Miller, E.K., Cohen, J.D.: An integrative theory of prefrontal cortex function. Annual review of neuroscience (2001)
59. OLMo Team, Ettinger, A., Bertsch, A., Kuehl, B., Graham, D., Heineman, D., Groeneveld, D., Brahman, F., Timbers, F., Ivison, H., et al.: Olmo 3. arXiv preprint arXiv:2512.13961 (2025)
60. Paiss, R., Ephrat, A., Tov, O., Zada, S., Mosseri, I., Irani, M., Dekel, T.: Teaching clip to count to ten. In: ICCV (2023)
61. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. IJCV (2017)
62. Pyatkin, V., Malik, S., Graf, V., Ivison, H., Huang, S., Dasigi, P., Lambert, N., Hajishirzi, H.: Generalizing verifiable instruction following. In: NeurIPS (2025)
63. Qi, P., Liu, Z., Zhou, X., Pang, T., Du, C., Lee, W.S., Lin, M.: Defeating the training-inference mismatch via fp16. arXiv preprint arXiv:2510.26788 (2025)
64. Qian, Y., Ye, H., Fauconnier, J.P., Grasch, P., Yang, Y., Gan, Z.: Mia-bench: Towards better instruction following evaluation of multimodal llms. In: ICLR (2025)
65. Qiao, R., Tan, Q., Yang, P., Wang, Y., Wang, X., Wan, E., Zhou, S., Dong, G., Zeng, Y., Xu, Y., Wang, J., Sun, C., Li, C., Zhang, H.: We-math 2.0: A versatile mathbook system for incentivizing visual mathematical reasoning. arXiv preprint arXiv:2508.10433 (2025)
66. Rogers, R.D., Monsell, S.: Costs of a predictible switch between simple cognitive tasks. Journal of experimental psychology: General (1995)
67. Sahu, P.P., Raut, A., Samant, J.S., Gorijala, M., Lakshminarayanan, V., Bhaskar, P.: Pop-vqa – privacy preserving, on-device, personalized visual question answering. In: WACV (2024)
68. Sanket Shah, Anand Mishra, N.Y., Talukdar, P.P.: Kvqa: Knowledge-aware visual question answering. In: AAAI (2019)
69. Sarch, G.H., Saha, S., Khandelwal, N., Jain, A., Tarr, M.J., Kumar, A., Fragkiadaki, K.: Grounded reinforcement learning for visual reasoning. In: NeurIPS (2025)
70. Schaul, T., Borsa, D., Modayil, J., Pascanu, R.: Ray interference: a source of plateaus in deep reinforcement learning. arXiv preprint arXiv:1904.11455 (2019)
71. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
72. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: ECCV (2022)

73. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: ICCV (2019)
74. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., et al.: DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
75. Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., Wu, C.: HybridFlow: A flexible and efficient RLHF framework. In: EuroSys (2025)
76. Shi, X., Lee, S.: Benchmarking out-of-distribution detection in visual question answering. In: WACV (2024)
77. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
78. Su, A., Wang, H., Ren, W., Lin, F., Chen, W.: Pixel reasoner: Incentivizing pixel-space reasoning via curiosity-driven reinforcement learning. In: NeurIPS (2025)
79. Tamarapalli, J.S., Grover, R., Pande, N., Yerramilli, S.: Countqa: How well do mllms count in the wild? arXiv preprint arXiv:2508.06585 (2025)
80. Tang, L., Kim, G., Zhao, X., Lake, T., Ding, W., Yin, F., Singhal, P., Wadhwa, M., Liu, Z.L., Sprague, Z., et al.: Chartmuseum: Testing visual reasoning capabilities of large vision-language models. In: NeurIPS (2025)
81. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
82. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
83. Team, G.R., Abeyruwan, S., Ainslie, J., Alayrac, J.B., Arenas, M.G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., Bohez, S., Bousmalis, K., Brohan, A., Buschmann, T., Byravan, A., Cabi, S., Caluwaerts, K., Casarini, F., Chang, O., Chen, J.E., Chen, X., Chiang, H.T.L., Choromanski, K., D'Ambrosio, D., Dasari, S., Davchev, T., Devin, C., Palo, N.D., Ding, T., Dostmohamed, A., Driess, D., Du, Y., Dwibedi, D., Elabd, M., Fantacci, C., Fong, C., Frey, E., Fu, C., Giustina, M., Gopalakrishnan, K., Graesser, L., Hasenclever, L., Heess, N., Hernaez, B., Herzog, A., Hofer, R.A., Humplik, J., Iscen, A., Jacob, M.G., Jain, D., Julian, R., Kalashnikov, D., Karagozler, M.E., Karp, S., Kew, C., Kirkland, J., Kirmani, S., Kuang, Y., Lampe, T., Laurens, A., Leal, I., Lee, A.X., Lee, T.W.E., Liang, J., Lin, Y., Maddineni, S., Majumdar, A., Michaely, A.H., Moreno, R., Neunert, M., Nori, F., Parada, C., Parisotto, E., Pastor, P., Pooley, A., Rao, K., Reymann, K., Sadigh, D., Saliceti, S., Sanketi, P., Sermanet, P., Shah, D., Sharma, M., Shea, K., Shu, C., Sindhwani, V., Singh, S., Soricut, R., Springenberg, J.T., Sterneck, R., Surdulescu, R., Tan, J., Tompson, J., Vanhoucke, V., Varley, J., Vesom, G., Vezzani, G., Vinyals, O., Wahid, A., Welker, S., Wohlhart, P., Xia, F., Xiao, T., Xie, A., Xie, J., Xu, P., Xu, S., Xu, Y., Xu, Z., Yang, Y., Yao, R., Yaroshenko, S., Yu, W., Yuan, W., Zhang, J., Zhang, T., Zhou, A., Zhou, Y.: Gemini robotics: Bringing ai into the physical world. arXiv preprint arXiv:2503.20020 (2025)
84. Teh, Y., Bapst, V., Czarnecki, W.M., Quan, J., Kirkpatrick, J., Hadsell, R., Heess, N., Pascanu, R.: Distral: Robust multitask reinforcement learning. In: NeurIPS (2017)

85. Tong, J., Tang, J., Li, H., Mou, Y., Zhang, M., Zhao, J., Wen, Y., Song, F., Zhan, J., Lu, Y., Tao, C., Guo, Z., Yu, J., Cheng, T., Xi, Z., Jiang, C., Yin, Z., Zheng, Y., Ge, W., Chen, G., Gui, T., Qiu, X., Zhang, Q., Huang, X.: Game-rl: Synthesizing multimodal verifiable game data to boost vlms' general reasoning. arXiv preprint arXiv:2505.13886 (2025)
86. Tong, P., Brown, E., Wu, P., Woo, S., Iyer, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In: NeurIPS (2024)
87. Wang, H., Qu, C., Huang, Z., Chu, W., Lin, F., Chen, W.: VL-Rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. In: NeurIPS (2025)
88. Wang, K., Pan, J., Shi, W., Lu, Z., Ren, H., Zhou, A., Zhan, M., Li, H.: Measuring multimodal mathematical reasoning with math-vision dataset. In: NeurIPS (2024)
89. Wang, P., Wu, Q., Shen, C., Dick, A., Van Den Hengel, A.: Fvqa: Fact-based visual question answering. TPAMI (2018)
90. Wang, Z., Xia, M., He, L., Chen, H., Liu, Y., Zhu, R., Liang, K., Wu, X., Liu, H., Malladi, S., et al.: Charxiv: Charting gaps in realistic chart understanding in multimodal llms. In: NeurIPS (2024)
91. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., Zhou, D.: Chain of thought prompting elicits reasoning in large language models. In: NeurIPS (2022)
92. Wiedmann, L., Zohar, O., Mahla, A., Wang, X., et al.: FineVision: Open data is all you need. arXiv preprint arXiv:2510.17269 (2025)
93. Wu, P., Zhang, Y., Diao, H., Li, B., Lu, L., Liu, Z.: Visual jigsaw post-training improves mllms. arXiv preprint arXiv:2509.25190 (2025)
94. Wu, Z., Wu, Z., Xu, F., Wang, Y., Sun, Q., Jia, C., Cheng, K., Ding, Z., Chen, L., Liang, P.P., Qiao, Y.: Os-atlas: A foundation action model for generalist gui agents. In: ICLR (2025)
95. xAI: Realworldqa: A benchmark for real-world spatial understanding. https://huggingface.co/datasets/xai-org/RealworldQA (2024), accessed: 2025-04-26
96. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. In: ICCV (2025)
97. Yang, J., Tan, R., Wu, Q., Zheng, R., Peng, B., Liang, Y., Gu, Y., Cai, M., Ye, S., Jang, J., Deng, Y., Liden, L., Gao, J.: Magma: A foundation model for multimodal ai agents. In: CVPR (2025)
98. Yang, Y., Patel, A., Deitke, M., Gupta, T., Weihs, L., Head, A., Yatskar, M., Callison-Burch, C., Krishna, R., Kembhavi, A., Clark, C.: Scaling text-rich image understanding via code-guided synthetic multimodal data generation. In: ACL (2025)
99. Yang, Y., Zhang, Z., Hou, Y., Li, Z., Liu, G., Payani, A., Ting, Y.S., Zheng, L.: Effective training data synthesis for improving mllm chart understanding. In: ICCV (2025)
100. Yao, H., Huang, J., Wu, W., Zhang, J., Wang, Y., Liu, S., Wang, Y., Song, Y., Feng, H., Shen, L., Tao, D.: Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. In: NeurIPS (2025)
101. Ying, K., Meng, F., Wang, J., Li, Z., Lin, H., Yang, Y., Zhang, H., Zhang, W., Lin, Y., Liu, S., et al.: Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. In: ICML (2024)
102. Yu, E., Lin, K., Zhao, L., Yin, J., Wei, Y., et al.: Perception-R1: Pioneering perception policy with reinforcement learning. In: NeurIPS (2025)

103. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., et al.: DAPO: An open-source LLM reinforcement learning system at scale. In: NeurIPS (2025)
104. Yu, W., Yang, Z., Ren, L., Li, L., Wang, J., Lin, K., Lin, C.C., Liu, Z., Wang, L., Wang, X.: Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. arXiv preprint arXiv:2408.00765 (2024)
105. Yue, X., Zheng, T., Ni, Y., Wang, Y., Zhang, K., Tong, S., Sun, Y., Yu, B., Zhang, G., Sun, H., et al.: Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In: ACL (2025)
106. Yue, Z., Lin, Z., Song, Y., Wang, W., Ren, S., et al.: MiMo-VL technical report. arXiv preprint arXiv:2506.03569 (2025)
107. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: CVPR (2019)
108. Zhang, C., Gao, F., Jia, B., Zhu, Y., Zhu, S.C.: Raven: A dataset for relational and analogical visual reasoning. In: CVPR (2019)
109. Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J.A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., et al.: Lmms-eval: Reality check on the evaluation of large multimodal models. In: Findings of NAACL (2025)
110. Zhang, K., Wu, K., Yang, Z., Li, B., Hu, K., Wang, B., Liu, Z., Li, X., Bing, L.: Openmmreasoner: Pushing the frontiers for multimodal reasoning with an open and general recipe. In: CVPR (2026)
111. Zhang, Y.F., Zhang, H., Tian, H., Fu, C., Zhang, S., Wu, J., Li, F., Wang, K., Wen, Q., Zhang, Z., et al.: Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In: ICLR (2025)
112. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. TMLR (2024)
113. Zheng, C., Liu, S., Li, M., Chen, X.H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., et al.: Group sequence policy optimization. arXiv preprint arXiv:2507.18071 (2025)
114. Zheng, L., Huang, Z., Xue, Z., Wang, X., An, B., Yan, S.: Agentstudio: A toolkit for building general virtual agents. In: ICLR (2025)
115. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images. In: CVPR (2016)