Curr-ReFT: Overcoming Training Bottlenecks in Small-scale Vision-Language Models via Curriculum Reinforcement Finetuning

Anonymous ACL submission

Abstract

State-of-the-art vision-language models (VLMs) require massive scaling that limits practical deployment. Small-scale VLMs offer a practical alternative but face out-of-domain (OOD) collapse when trained with traditional supervised fine-tuning (SFT). Our experiments reveal that RL-based post-training can mitigate this OOD degradation, but faces a critical sparse reward dilemma in complex visual reasoning tasks. To this end, we propose Curriculum Reinforcement Finetuning (Curr-ReFT), comprising two sequential stages: (1) Structured Curriculum Reinforcement Learning, which progressively evolves task formats and reward functions to match models' growing capabilities; and (2) Rejected Sampling-based Self-improvement, which maintains the fundamental capabilities of VLMs through selective learning from high-quality examples. Extensive experiments demonstrate that Curr-ReFT achieves state-of-the-art performance across various visual tasks in both in- and out-of-domain settings and benchmarks. Notably, our Curr-ReFT-7B achieves performance comparable to 26B-scale models on multiple benchmarks.

1 Introduction

Recent advances in multimodal understanding have led to remarkable vision-language models (VLMs), exemplified by OpenAI (Arrieta et al., 2025; Jaech et al., 2024; Wainwright and Lowe, 2023), InterVL (Chen et al., 2024b; Wang et al., 2022), and QWen (Wang et al., 2024; Yang et al., 2024) series. However, these achievements predominantly rely on massive model scaling (>32B parameters), creating substantial deployment barriers in resource-constrained environments. This limitation motivates the exploration of efficient training paradigms for small-scale VLMs (<10B parameters).

While supervised fine-tuning (SFT) with high-quality annotated data (Bai et al., 2022; Ziegler et al., 2019) is the dominant training approach for VLMs, it poses a critical challenge for small-scale VLMs: **generalization collapse** (Abbas et al., 2025; Srivastava et al., 2025; Yu et al., 2025). As evidenced in Fig. 1(a), SFT-trained models consistently outperform base models on in-domain data across detection, classification, and multimodal math tasks. However, these gains are accompanied by significant performance degradation on out-of-domain (OOD) distributions, underscoring the challenge of "OOD degradation". This phenomenon aligns with recent theoretical findings

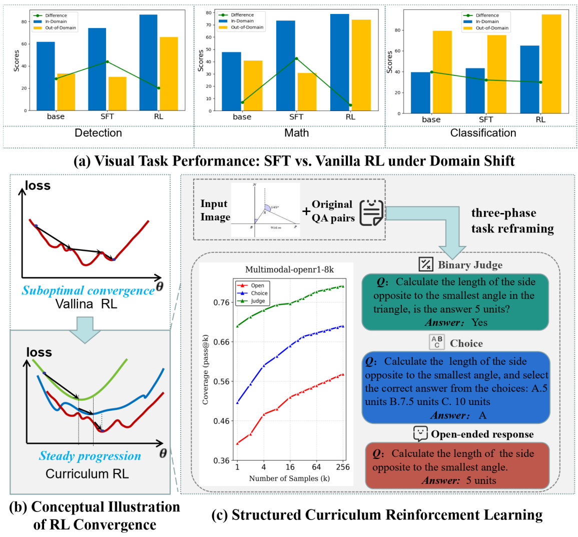


Figure 1: (a) Visual Task Performance: SFT vs. Vanilla RL under in- and out-of-domain settings. (b) Illustration of Vallina RL vs Curriculum RL. The "Training Bottleneck" in small-scale VLMs: suboptimal convergence when facing complex visual reasoning tasks. Our Curriculum RL ensures steady progression of model training through Phased Task Reframing and Hierarchical Reward Design. (c) Structured Curriculum Reinforcement Learning. Curr-RL systematically reformulates input questions into three progressively complex formats. Using multimodal math (OpenR1-8k) as the test case, the pass@k(Cheng et al., 2024) curves reveal a fundamental trade-off between solution space constraints and reasoning complexity.

(Fu et al., 2023; Srivastava et al., 2025) attributing OOD degradation to SFT's inherent bias toward pattern memorization rather than systematic reasoning (Yu et al., 2025).

DeepSeek R1-Zero's (Guo et al., 2025) success with Group Relative Policy Optimization (GRPO) suggests a promising direction for enhancing reasoning through comparative response evaluation. Motivated by these advances, we investigate *whether RL-based post-training can similarly enhance OOD generalization in small-scale vision-language models.* Comprehensive experiments reveal a consistent pattern: while SFT suffers significant OOD degradation, RL-based methods maintain robust generalization across diverse visual tasks (Fig. 1 (a)).

Although RL effectively mitigates OOD challenges, it encounters **'Training Bottleneck'** in visual reasoning tasks, characterized by minimal policy updates, premature convergence to suboptimal strategies, and repetitive generation of low-quality responses (Fig. 1 (b)). This bottleneck arises from **sparse reward** (Xi et al., 2024; Tec et al., 2025; Wei et al., 2023)—***tasks with complex solution spaces provide rarely positive feedback, leading to insufficient learning and suboptimal convergence.***

To address this bottleneck, we propose a Structured Curriculum Reinforcement Learning (Curr-RL) paradigm that progressively evolves task formats to match models' growing capabilities, inspired by curriculum learning (Kong et al., 2021; Pentina et al., 2015; Lin et al., 2023; Ryu et al., 2024). ***Our key insight is that the sparse reward dilemma mainly arises from the vast solution space, hindering the exploration of correct paths, particularly in early-stage training.*** (Lin et al., 2023). As illustrated in Fig. 1 (c), Curr-RL employs a three-phase task reframing and hierarchical reward design, enabling smooth transitions from structured to open-ended formats. This progression begins with binary decisions that reconstruct complex visual reasoning into true/false questions, reducing the solution space dimension for more dense rewards. It then progresses to choice selection formats that introduce partially open elements, and culminates in open-ended generation, developing robust vision-language associations before confronting sparse reward scenarios.

While Curriculum RL effectively boosts domain-specific visual reasoning, it often compromises general-purpose language capabilities (**e.g.,** commonsense and scientific reasoning), a known trade-off in RL fine-tuning (Zhang et al., 2023; Pan et al., 2024; Hafez and Erekmen, 2024). To address this issue, we introduce a rejection sampling-based self-improvement mechanism that preserves general knowledge. Built upon the Crescent framework (Team et al., 2025; Guo et al., 2025), our method employs an LLM-as-judge (**e.g.,** GPT-4o (Wainwright and Lowe, 2023)) to compare the RL-trained model's responses with reference answers and retain the higher-quality responses.

To this end, we propose Curriculum Reinforcement Finetuning (**Curr-ReFT**). Curr-ReFT comprises two sequential stages: Structured Curriculum Reinforcement Learning and Rejected Sample-based Self-improvement. Extensive experiments demonstrate Curr-ReFT's state-of-the-art performance on both in- and out-of-domain visual tasks and abundant public benchmarks, with our enhanced small-scale VLMs matching the capabilities of much larger counterparts.

Contributions Summary. (1) **Theoretical Insight:** We demonstrate that rule-based reinforcement learning enhances OOD generalization in visual tasks without extra data; (2) **Curr-ReFT Framework:** An adaptable two-stage post-training paradigm that strengthens visual reasoning while preserving fundamental language capabilities; (3) **Curriculum Dataset:** A newly constructed 12k-example dataset spanning detection, classification, and multimodal math; (4) **Empirical Results:** Extensive experiments demonstrate Curr-ReFT's superior performance across multiple benchmarks.

2 Related Work

2.1 Reasoning Vision-language models

Recent advancements in multimodal models have evolved from LLaVA's (Liu et al., 2023) projection-based approach to Qwen-VL (Bai et al., 2023; Wang et al., 2024; Yang et al., 2024) and InternVL (Chen et al., 2024a,b; Luo et al., 2024) series further advancing visual instruction tuning and architectural efficiency. Concurrently, reasoning-focused methods have progressed from Monte Carlo Tree Search techniques (Browne et al., 2012; Yao et al., 2023) to process-supervised learning (Lu et al., 2024). OpenAI-O1 (Wainwright and Lowe, 2023) established the RL+SFT paradigm, while DeepSeek-R1-Zero's GRPO (Guo et al., 2025) demonstrated superior reasoning through group-wise response comparisons without auxiliary networks (Schulman et al., 2017). Despite these ad-

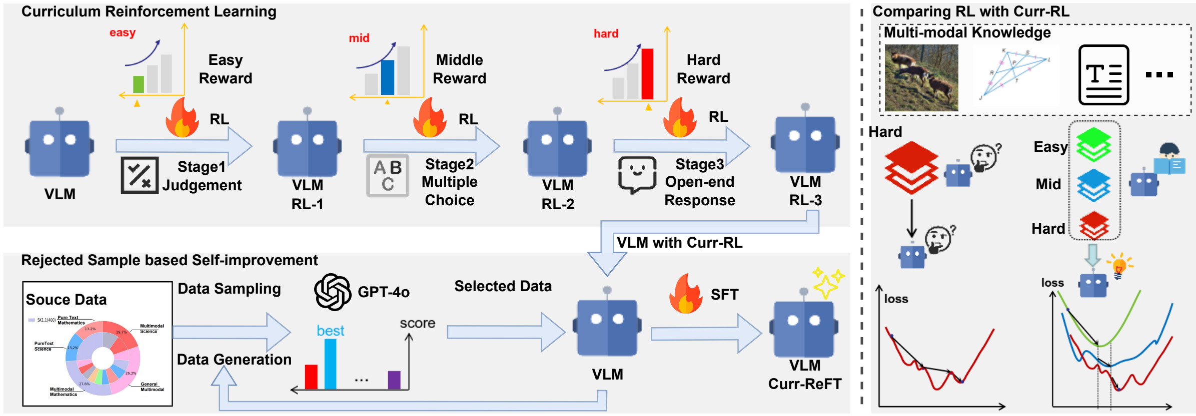


Figure 2: **Overall framework of the proposed Curr-ReFT post-training paradigm.** Curr-ReFT comprises two sequential stages: (1) Curriculum Reinforcement Learning that progressively increases task complexity with aligned reward mechanisms, and (2) Rejected Sample based Self-improvement that maintains fundamental capabilities (Best viewed in color).

vances, current research primarily targets math and coding tasks (Liu et al., 2024a), leaving the intersection of visual perception and reasoning largely unexplored. Our Curr-ReFT framework addresses this through multi-stage RL training.

3 Method

In this section, we elaborate **Curr-ReFT**, comprising two sequential training stages: Curriculum Reinforcement Learning (Sec. 3.2), which achieves task progression training through three stages of reward mechanisms, and Rejected Sample based Self-improvement (Sec. 3.3), which preserves fundamental capabilities via quality-guided learning. The overall framework is illustrated in Fig. 2.

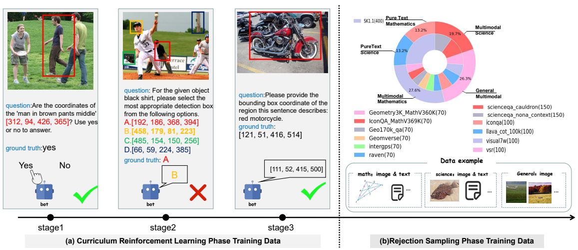


Figure 3: **Illustration of training data organization.** (a) Examples of 3-stage progressive response formats in Curriculum Reinforcement Learning. (b) Data source in Reject-sampling SFT phase (detailed Reject-sampling pipeline in Sec. 3.3).

3.1 Preliminary

Reinforcement Learning with GRPO

DeepSeek R1-Zero (Guo et al., 2025) introduces the GRPO framework, eliminating dependence on additional critic networks (PPO-based methods(Schulman et al., 2017)). Specifically, GRPO considers the relative performance of responses rather than absolute reward values. For a given input query q. The framework generates N distinct responses $\{o_1, o_2, ..., o_N\}$ from the current policy π_θ and evaluates through group-wise comparison:

$$A_i = \frac{r_i - \mathrm{mean}(\{r_1, \ldots, r_N\})}{\mathrm{std}(\{r_1, \ldots, r_N\})} \tag{1}$$

where A_i represents the normalized relative quality of the i-th response within its group.

3.2 Structured Curriculum Reinforcement Learning

The Structured Curriculum Reinforcement Learning employs a three-phase dynamic adjustment on task formats and reward functions to address RL's sparse reward issue. We will elaborate Binary Decision Learning, Multiple Choice Learning, and Open-ended Response on the task formats and reward designs in Sec. 3.2.1, Sec. 3.2.2, and Sec. 3.2.3, respectively.

3.2.1 Stage 1: Binary Decision Learning

In the initial stage of reinforcement learning, we adopt binary decision questions as the simplest form of task format, as shown in Fig. 3 (a), which significantly reduces the output freedom to binary choices, making it easier to learn basic visual understanding and reasoning patterns. Models are explicitly prompted to answer with 'yes' or 'no.' The reward function for this stage is as follows:

$$\mathbf{R}_{\mathbf{Binary}}(\mathbf{o}_{std}, \mathbf{o}_{gt}) = \begin{cases} 1, & \text{if } \mathbf{o}_{std} = \mathbf{o}_{gt} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\mathbf{o}_{std}$ represents the model's binary response and $\mathbf{o}_{gt}$ is the ground truth answer.

3.2.2 Stage 2: Multiple Choice Learning

The second stage introduces choice questions, which require more sophisticated decision-making while maintaining structured response formats (as displayed in Fig. 3 (a). We design different reward mechanisms for single-choice and multiple-choice scenarios to provide appropriate learning signals. For single-choice questions, we maintain a binary reward structure:

$$\mathbf{R}_s(\mathbf{o}_{std}, \mathbf{o}_{gt}) = \begin{cases} 1, & \mathbf{o}_{std} = \mathbf{o}_{gt} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

For multiple-choice questions, we introduce a more nuanced reward function that considers partial correctness:

$$\mathbf{R}_m(\mathbf{o}_{std}, \mathbf{o}_{gt}) = \begin{cases} 1, & \mathbf{o}_{std} = \mathbf{o}_{gt} \\ 0.2, & \mathbf{o}_{std} \subset \mathbf{o}_{gt}, |\mathbf{o}_{std}| > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $\mathbf{o}_{std}$ represents the model's selected options and $\mathbf{o}_{gt}$ is the set of correct options. This graduated reward structure encourages the model to identify correct options while maintaining the incentive for complete answers.

3.2.3 Stage 3: Open-ended Response Learning

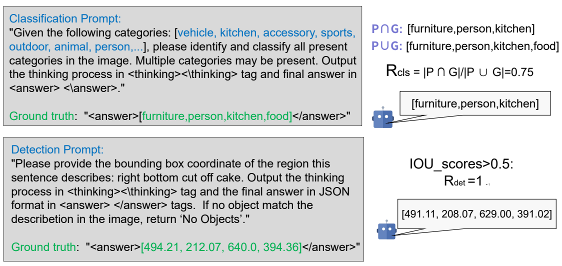


Figure 4: Verifiable Reward for visual tasks in the Open-ended Response stage. We have listed the detection and classification prompt with Verifiable Reward calculation examples.

Inspired by DeepSeek-R1's success in reasoning tasks, we extend its RL approach to visual domains. Unlike math or code tasks with clear ground truth, visual tasks require tailored reward functions. We design verifiable, task-specific rewards for open-ended multimodal RL.

Category Overlap Reward for Visual Classification For classification, we propose a Category Overlap Reward, computed as the intersection-over-union (IoU) between predicted and ground-truth categories. This continuous reward offers proportional credit for partial correctness, providing richer feedback than binary matching. Let the predicted categories be $P = c_1, c_2, ..., c_m$ and ground-truth categories $G = g_1, g_2, ..., g_n$, where c_i and g_j denote individual category labels. The reward is calculated based on their set intersection and union:

$$\mathbf{R}_{acc\_cls} = \frac{|P \cap G|}{|P \cup G|} = \frac{|\{c_i | c_i \in P \text{ and } c_i \in G\}|}{|\{c_1, ..., c_m\} \cup \{g_1, ..., g_n\}|}, \tag{5}$$

where $|P \cap G|$ represents the number of correctly predicted categories, and $|P \cup G|$ represents the total number of unique categories in both sets combined. This reward mechanism provides a continuous value in [0,1], better reflecting partial correctness in multi-label scenarios compared to binary rewards. The classification reward R_{cls} combines accuracy and format compliance.

IOU rewards for Visual Detection For object detection tasks, we design a comprehensive reward function that evaluates both localization accuracy. The reward mechanism considers three key aspects: spatial accuracy, prediction reliability, and response format compliance.

Given a set of predicted bounding boxes $B_{student} = \{b_1, b_2, ..., b_n\}$ with corresponding confidence scores $f = \{f_1, f_2, ..., f_n\}$, and ground truth boxes $B_{gt} = \{b_1^{gt}, b_2^{gt}, ..., b_m^{gt}\}$, we first establish box-level correspondences through IoU matching. By applying a threshold τ, we filter out low-quality matches where $iou_i < \tau$. The localization accuracy reward R_{loc} is then computed as the mean IoU of the remaining valid matches:

$$\mathbf{R}_{Iou} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} iou_i, \quad \mathcal{V} = \{i | iou_i \geq \tau\} \tag{6}$$

where $\mathcal{V}$ denotes the set of valid matches and $|\mathcal{V}|$ represents the number of valid matches. To encourage accurate object localization, we further discretize the IoU-based reward using a threshold of 0.5:

$$\mathbf{R}_{acc\_det} = \begin{cases} 1, & \text{if } \mathbf{R}_{Iou} > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

The final detection reward $\mathbf{R}_{det}$ combines both localization accuracy and format compliance:

$$\mathbf{R}_{det} = \mathbf{R}_{acc\_det} + \mathbf{R}_{format} \tag{8}$$

where $\mathbf{R}_{acc\_det}$ evaluates spatial localization accuracy and R_{format} verifies response format compliance.

3.3 Rejected Sample based Self-improvement

To prevent degradation in general domain, we propose a Rejected Sample-based Self-improvement mechanism. This framework comprises three stages: (1) general-domain data construction; (2) automatic refinement by GPT-4-O–guided selection between model and reference responses, enhancing data quality via self-improvement; and (3) supervised fine-tuning on the curated dataset to reinforce general vision-language competence. Specifically, step (1) is detailed in Sec. 3.3.1, while steps (2) and (3) are elaborated in Sec. 3.3.2.

3.3.1 General-domain data construction

The data preparation process involves systematic sampling from a comprehensive dataset. Utilizing GPT-4-O as the reward model, we evaluate generated responses against multiple criteria: accuracy, logical consistency, format compliance, and linguistic fluency. Responses are quantitatively assessed on a 0-100 scale, with those surpassing a threshold of 85 being integrated into the enhanced dataset alongside their corresponding queries. The resultant curated dataset encompasses 1,520 high-quality samples (12.7% selection rate) across science, general knowledge and math (Fig. 3).

3.3.2 Self-Improvement Training

Following dataset construction, we refine the data by selecting high-quality answers from either the original references or model-generated responses, as determined by LLM-as-Judge (GPT-4O(Wainwright and Lowe, 2023)). As illustrated in Algorithm 1, for each query in the dataset, the model generates multiple candidate responses, which are scored by LLM-as-Judge. The highest-scoring response—regardless of whether it is model-generated or a reference—is retained if it surpasses a quality threshold. The resulting refined dataset is then used to conduct further fine-tuning. This self-improvement stage reinforces general-domain competencies by enabling the model to learn from its own superior outputs while maintaining robust cross-domain capabilities.

Algorithm 1 Rejected Sample based Self-improvement Self-Improvement Algorithm

Require: Dataset D, Generative Model G, Reward Model R, Number of samples per input N, Threshold score $\tau = 85$

Ensure: Enhanced dataset D_{new}

1: Initialize an empty enhanced dataset $D_{\text{new}} \leftarrow \emptyset$
2: **for** each question $q \in D$ **do**
3:     Generate N responses using the generative model G: $\{r_1, r_2, \ldots, r_N\} = G(q)$
4:     Score each response using the reward model R: $\{s_1, s_2, \ldots, s_N\} = R(\{r_1, r_2, \ldots, r_N\})$
5:     Find the index of the highest-scored response: $i^* = \arg\max_i s_i$
6:     **if** $s_{i^*} > \tau$ **then**
7:         Add the highest-scored response to: $D_{\text{new}} \leftarrow D_{\text{new}} \cup \{(q, r_{i^*})\}$
8:     **end if**
9: **end for**
10: Return the enhanced dataset D_{new}

4 Experiments

Aiming to answer the following questions, we conduct extensive experiments and test on abundant benchmarks:

- **RQ1:** How does RL perform compared to traditional SFT in standard CV tasks?
- **RQ2:** How do models trained with Curr-ReFT perform relative to mainstream VLMs?
- **RQ3:** How do curriculum strategies like order rearrangement impact Curriculum RL performance? How do curriculum learning and rejection sampling contribute to performance in general and visual tasks, respectively?
- **RQ4:** Is Curr-ReFT effective across different backbone models and model sizes?

4.1 Experiment Settings

4.1.1 Datasets and Metrics

Datasets. We built an evaluation framework across three tasks: **visual detection**, **visual classification**, and **multimodal mathematical reasoning**—each with 4,000 training, 1,000 in-domain test and 1,000 out-domain samples. For training and in-domain test sets, detection and classification data comes from RefCOCO (Yu et al., 2016) and RefCOCOg (Mao et al., 2016), while math data use Math360K (Shi et al., 2024) and Geo170K (Gao et al., 2023). All training samples are reformatted into binary decision, choice, and open-ended formats for three-stage Curriculum RL. Out-domain evaluation includes RefGTA (Tanaka et al., 2019) for detection, Pascal-VOC (Everingham et al., 2010) for classification, and CLEVR-70k-Counting for math.

Metrics. We use accuracy as the unified evaluation metric, defined as correct predictions over total test samples. For detection, a prediction is correct if the IoU between predicted and ground truth boxes exceeds 0.5. In classification, predictions matching ground truth labels are considered correct.

4.1.2 Benchmarks

To give a reasonable result, we evaluate our trained models on the following authoritative benchmarks: MathVista (Lu et al., 2023), MATH (Hendrycks et al., 2021), AI2D (Hiippala et al., 2021), MMVet (Yu et al., 2023), MMBench (Liu et al., 2024b), OCRBench (Liu et al., 2024c), and LLaVABench (Liu et al., 2023), covering a wide range of mathematical reasoning, scientific understanding, visual perception, and multimodal generalization tasks.

4.1.3 Training Details

All experiments use NVIDIA A800 GPUs. We primarily train Qwen2.5-VL-3B on 8 GPUs (batch size=8), with scaling tests using Qwen2.5-VL-7B across 16 GPUs. The hyperparameters are set as follows: (1) Learning rates: 2e-5 for RL (GRPO) training, 2e-7 for rejection sampling phase, and 1e-6 for SFT experiments. (2) Maximum pixel size: 401,408. (3) GRPO training steps: 2,500.

The training methods are described as follows: (1)'**+SFT**' denotes supervised fine-tuning using 12k multimodal data on open-ended response formats. (2)'**+ReFT**' denotes rejection-sampling-based SFT; the data and sampling strategy are detailed in Sec. 3.3. (3)'**+RL**' refers to direct GRPO training using the same 12k samples, with reward functions aligned with Stage 3 of Curr-RL in Sec. 3.2.3. (4)'**+Curr-RL**' involves the three-stage curriculum RL training as detailed in Sec. 3.2. (5)'**+RL-ReFT**' and '**+Curr-ReFT**' indicate applying rejection-sampling-based fine-tuning after RL (GRPO) or curriculum RL, respectively.

4.2 Main Results

4.2.1 Generalization Verification of RL (RQ1)

Tab. 1 summarizes the in- and out-of-domain performance of Qwen2.5-VL-3B under SFT, RL, and Curriculum RL paradigms. Fig. 6 provides qualitative comparisons between SFT and our Curr-RL method. Key observations are as follows: **(1)** While SFT improves in-domain accuracy, it consistently fails on out-of-domain tasks, highlighting limited generalization of SFT paradigm. **(2)** RL-based methods exhibit stronger generalization, with Curr-RL achieving the best overall performance. In particular, it enhances both mathematical reasoning and visual perception. **(3)** Qualitatively, Curr-RL generates more accurate localizations and comprehensive explanations across diverse out-of-domain settings.

Methods	In-domain			Out-domain		
	Det	Math	Cls	Det	Math	Cls
Base	61.8	71.3	39.6	55.3	40.8	79.3
+SFT	75.2	73.5	50.2	52.3	30.8	77.2
+RL	88.3	78.8	62.9	64.2	74.1	94.7
+Curr-RL	**90.6**	**82.8**	**66.8**	**67.1**	**78.4**	**96.6**

Table 1: Performance Comparison: In/Out-domain Performance (%). Base model is chosen as the Qwen2.5-VL-3B. Notably, 'Det' and 'Cls' denote detection and classification.

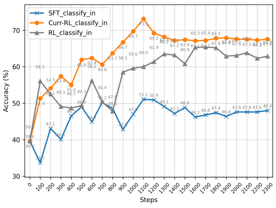
Fig. 5 further presents the training dynamics of SFT, RL, and Curr-RL. **(1)** Curr-RL demonstrates more stable training, faster convergence than RL, and achieves the highest accuracy after convergence. **(2)** Curr-ReFT achieves larger gains over RL in visual classification, likely due to the greater reliance on semantic reasoning. In contrast, detection tasks focus on spatial perception. Curriculum mechanisms are particularly effective for semantically demanding tasks.

4.2.2 Performance Comparation (RQ2)

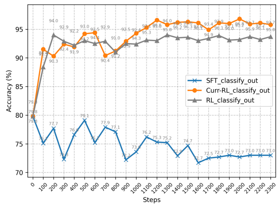
We report results on visual datasets (Tab. 2) and public benchmarks (Tab. 3), using Curr-ReFT-3B/7B initialized from Qwen2.5-VL-3B/7B. Our

Methods	**Math**		**Detection**		**Classification**	
	In	**Out**	**In**	**Out**	**In**	**Out**
Qwen2.5-VL-3B	71.3	40.8	61.8	55.3	39.6	79.3
InternVL2_5-4B	69.4	36.3	60.2	54.5	41.5	78.9
Qwen2.5-VL-7B	77.9	54.6	76.7	63.6	62.5	81.3
InternVL2_5-8B	76.3	52.1	67.1	59.7	61.1	82.9
InterVL2-26B	83.7	77.4	78.9	68.3	68.1	91.5
LLaVA-32B	83.4	78.7	81.2	65.4	69.6	93.4
Qwen2.5-VL-3B†	73.5	30.8	75.2	52.3	50.2	77.2
InternVL2_5-4B†	72.1	31.8	64.1	52.8	48.5	73.3
InternVL2_5-8B†	80.9	46.9	89.7	41.2	68.9	79.2
Curr-ReFT-3B	82.3	73.7	89.8	65.6	71.5	95.2
Curr-ReFT-7B	**85.3**	**81.5**	**92.2**	**69.5**	**73.1**	**98.7**

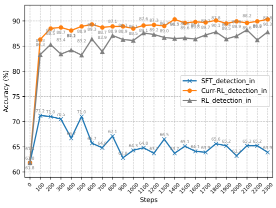
Table 2: **Performance comparison on visual tasks.**'In' denotes in-domain while 'out' represents out-of-domain testing. SFT† results are shown with the slash. Boldface and underlines indicate the best and second-best results.



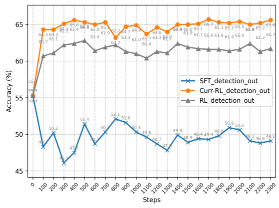
(a) In-domain Classification



(b) Out-domain Classification



(c) In-domain Detection



(d) Out-domain Detection

Figure 5: Performance Dynamics: SFT vs. RL vs. Curriculum RL on In-domain and Out-of-domain Tasks.

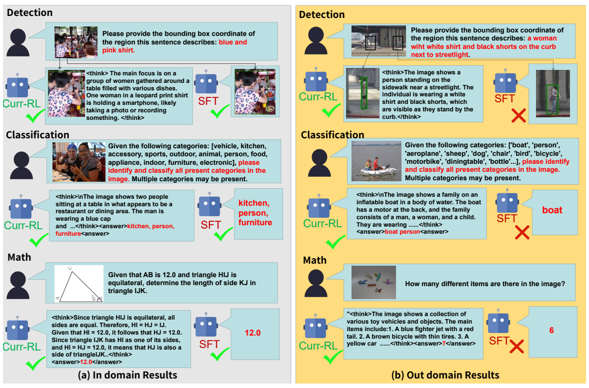


Figure 6: Qualitative comparison between our method and SFT baseline. Thinking process significantly improves reasoning ability.

results reveal the following:

Visual Tasks Curr-ReFT shows strong improvements under OOD conditions. For math reasoning, Curr-ReFT-7B achieves 81.5% OOD accuracy, outperforming its base model by +34.6%, and achieves 69.5% and 98.7% on OOD detection and classification, respectively—exceeding LLaVA-32B by a large margin with only 25% of its parameters.

Public Benchmarks Curr-ReFT-7B performs competitively with much larger models (26B/32B) on most benchmarks. On LLaVABench, it reaches 83.6% (Relative), 69.7% (VLM), and 84.5% (GPT-4o), surpassing LLaVA-32B (82.2%) and approaching GPT-4o's level. This demonstrates the effectiveness of our rejection-based self-improvement in enhancing general vision-language capacity.

In addition, we observe two noteworthy findings: **(1)** OOD classification outperforms in-domain by +25.6%, likely due to clearer semantics in OOD test data versus more ambiguous in-domain labels. **(2)** SFT remains strong on structured tasks like detection, but underperforms in reasoning math tasks.

4.2.3 Ablation Study (RQ3)

Tab. 4(a) and Tab. 4(b) report ablation results on visual tasks and benchmarks, respectively. We compare Curr-ReFT-3B with the following variants: (1) **+Curr-RL**, without reject-sampling SFT; (2) **+RL**, without curriculum learning. To examine task sensitivity of curriculum learning, we compare **+RL**$_{\textbf{Judge}}$ and **+RL**$_{\textbf{Choice}}$ (purely judgment or choice formats), **+Curr-RL**$^{\textbf{Reverse}}$ (reverse curriculum, starting with open-ended response with finally judgment format **+Curr-RL**$^{\textbf{Mix}}$ (a mixed strategy that uses an equal proportion of the three formats). Our ablation study reveals four key insights:

- Curr-RL consistently outperforms standard RL. Progressive order yields better results than randomized (Mix) or reversed (Reverse) curricula.
- Reject-sampling improves language-centric benchmarks but compromises visual grounding,

Model	AI2D	MMVet	MMBencl	LLaVABench			MathVist	MATH	OCRBench
				Relative	VLM	GPT4			
Qwen2.5-VL-3B	74.35	39.04	63.00	67.30	56.30	80.40	52.00	38.60	59.70
InternVL2_5-4B	75.84	<u>42.48</u>	73.25	59.50	54.10	78.20	56.00	40.10	68.10
Qwen2-VL-7B	79.10	39.40	70.01	60.20	51.20	74.50	49.40	30.70	65.50
Qwen2.5-VL-7B	78.01	47.69	75.10	<u>81.20</u>	67.10	81.40	<u>63.80</u>	43.40	67.60
InternVL2_5-8B	65.42	34.17	52.30	72.10	63.70	80.10	50.10	42.10	59.80
InterVL2-26B	78.89	42.11	72.80	78.10	74.11	81.00	51.40	37.10	62.80
LLaVA-32b	78.90	45.20	<u>77.85</u>	80.21	**75.30**	82.20	57.10	40.10	<u>70.20</u>
Curr-ReFT-3B	<u>79.66</u>	39.95	69.40	73.80	68.10	**85.60**	57.90	<u>45.80</u>	62.30
Curr-ReFT-7B	**83.16**	**49.95**	**80.12**	**83.60**	<u>69.70</u>	<u>84.50</u>	**65.80**	**56.60**	**72.70**

Table 3: **Performance comparison against mainstream VLMs on public benchmarks.** Background colors denote benchmark categories: yellow for science (AI2D), cyan for general vision-language understanding (MMVet, MMBench, LLaVABench), green for math-related tasks (MathVista, MATH), and red for OCR (OCRBench). Boldface and underlines indicate the best and second-best results, respectively.

Method	Math		Detection		Classification	
	In	Out	In	Out	In	Out
Base	71.3	17.8	31.8	22.3	39.6	79.8
+SFT	73.5	30.8	75.2	52.3	50.2	77.2
+RL$_{\text{Judge}}$	76.2	71.9	89.4	65.1	63.0	94.1
+RL$_{\text{Choice}}$	77.1	73.2	88.1	63.5	64.9	94.5
+RL	78.8	74.1	88.3	64.2	62.9	93.8
+Curr-RL$^{\text{Reverse}}$	72.4	72.4	80.2	66.1	60.9	94.3
+Curr-RL$^{\text{Mix}}$	77.8	75.4	85.6	67.1	61.8	94.6
+Curr-RL	**82.8**	**78.4**	**90.6**	**67.1**	**66.8**	**96.6**
+ReFT	69.5	52.5	25.7	49.7	39.2	72.4
+RL-ReFT	77.1	70.3	84.3	62.1	54.9	94.5
+Curr-ReFT	80.3	73.7	89.8	65.6	65.4	92.2

(a) **Ablation study results on visual tasks** 'Base' denotes Qwen2.5-VL-3B model. Notably, color highlighting indicates different training strategies: Vision-specific SFT results (green), RL training scheme (blue), General-domain Reject-sampling SFT (yellow), and proposed Curr-ReFT (gray). Details of '**+RL$_{\text{Judge}}$**', '**+RL$_{\text{Choice}}$**', '**+Curr-RL$^{\text{Reverse}}$**', and '**+Curr-RL$^{\text{Mix}}$**' are provided in Sec. 4.2.3.

Method	AI2D	MMVet	MathV	OCR	MM	LLaVA	
						VLM	GPT4
Base	74.35	39.04	52.00	59.70	63.32	63.30	80.40
+SFT	75.45	32.02	53.60	58.90	63.65	59.30	84.10
+RL	76.46	36.28	55.30	<u>60.90</u>	66.34	<u>64.70</u>	83.00
+Curr-RL	77.36	36.74	<u>56.30</u>	59.40	<u>69.02</u>	64.10	<u>85.20</u>
+ReFT	<u>78.65</u>	<u>38.95</u>	52.80	59.50	67.90	66.10	84.20
+Curr-ReFT	**79.66**	**39.95**	**57.90**	**62.30**	**69.78**	**68.10**	**85.60**

(b) **Ablation study on standard benchmarks. Color coding follows the same scheme as in Tab. 4a.**

Table 4: **Ablation Study on major components.**

suggesting that general-domain data weakens alignment with fine-grained visual cues.

- SFT alone is insufficient for generalization under distribution shift. Despite strong in-domain performance, it performs poorly OOD, especially on visual reasoning. The absence of feedback limits its ability to generalize beyond training data.
- Combining Curr-RL with ReFT yields complementary benefits. Curr-ReFT unifies curriculum-driven progression and rejection-based generalization, yielding balanced gains in perception and reasoning.

4.2.4 Scaling Analysis (RQ4)

To examine the scaling effectiveness of our Curr-ReFT framework, we conduct extensive experiments on base models of varying sizes and types. The results in Tab. 2, Tab. 3 and Tab. 5 indicate that the effectiveness of Curr-ReFT scales effectively with model size:

- Compared to the Curr-ReFT-3B, Curr-ReFT-7B shows consistent improvements: Higher performance on visual tasks (detection: 89.8% → 92.2%, classification: 71.5% → 73.1%) and better generalization on benchmarks (MMVet: 39.95% → 49.95%, MathVista: 57.90% → 65.80%).
- Curr-ReFT consistently improves performance across different models (InternVL2_5-4B/8B, Qwen2-VL-7B, and Qwen2.5-VL-3B/7B), with gains becoming more pronounced as model size increases (Detection (InternVL2_5-4B: 54.5% → 61.2% , ↑6.7; InternVL2_5-8B: 59.7% → 70.4% , ↑10.7)).

Method	Math		Detection		Classification	
	In	Out	In	Out	In	Out
InternVL2_5-4B	69.4	36.3	60.2	54.5	41.5	78.9
+Curr-ReFT	**76.8** ↑7.4	**46.7** ↑10.4	**68.4** ↑8.2	**61.2** ↑6.7	**50.2** ↑8.7	**87.3** ↑8.4
InternVL2_5-8B	76.3	52.1	67.1	59.7	61.1	82.9
+Curr-ReFT	**83.1** ↑6.8	**60.2** ↑8.1	**76.8** ↑9.7	**70.4** ↑10.7	**67.2** ↑6.1	**90.1** ↑7.2
Qwen2-VL-7B	77.9	54.6	76.7	63.6	62.5	81.3
+Curr-ReFT	**85.3** ↑7.4	**62.6** ↑8.0	**84.8** ↑8.1	**69.5** ↑5.9	**69.6** ↑7.1	**87.5** ↑6.2

Table 5: **Scaling Up Experiment on visual dataset:** We evaluate the scalability of the proposed *Curr-ReFT* framework on various vision-language base models. Red ↑ indicates the relative gain.

5 Conclusion

In this paper, we introduce Curr-ReFT, a novel two-stage post-training paradigm that balances domain-specific visual reasoning and general vision-language capabilities. We provide theoretical insights that reinforcement learning improves both reasoning and out-of-domain visual task generalization. We also release a 12k-example curriculum benchmark spanning visual detection, classification, and multimodal math. Curr-ReFT achieves SOTA on abundant benchmarks, with +5.2% avg. gain in OOD visual tasks.

Limitations and Future Works

While Curr-ReFT effectively improves reasoning and generalization in small-scale VLMs, several potential limitations merit consideration. First, the constrained task formats used in early curriculum stages (e.g., binary or multiple-choice) may bias the model toward generating shorter or less diverse responses in open-ended tasks. Second, although the progressive task transition facilitates stable learning, it may also risk catastrophic forgetting of early-stage skills if not complemented by explicit retention mechanisms. Third, our three-stage curriculum is manually designed, which may limit scalability across domains or modalities.

We partially address the second concern through a rejection-sampling-based self-improvement that reinforces general capabilities. Nonetheless, further exploration of automated curriculum scheduling and lifelong learning strategies remains a promising direction.

References

Momin Abbas, Muneeza Azmat, Raya Horesh, and Mikhail Yurochkin. 2025. Out-of-distribution detection using synthetic data generation. *arXiv preprint arXiv:2502.03323*.

Aitor Arrieta, Miriam Ugarte, Pablo Valle, José Antonio Parejo, and Sergio Segura. 2025. Early external safety testing of openai's o3-mini: Insights from the pre-deployment evaluation. *arXiv preprint arXiv:2501.17749*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. 2012. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Zelei Cheng, Xian Wu, Jiahao Yu, Sabrina Yang, Gang Wang, and Xinyu Xing. 2024. Rice: Breaking through the training bottlenecks of reinforcement learning with explanation. *arXiv preprint arXiv:2405.03064*.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *International Conference on Machine Learning*, pages 10421–10430. PMLR.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and 1 others. 2023. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Muhammad Burhan Hafez and Kerim Erekmen. 2024. Continual deep reinforcement learning with task-agnostic policy distillation. *Scientific Reports*, 14(1):31661.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. 2021. Ai2d-rst: a multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55:661–688.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Yajing Kong, Liu Liu, Jun Wang, and Dacheng Tao. 2021. Adaptive curriculum learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5067–5076.

Fanqi Lin, Shiyu Huang, Tim Pearce, Wenze Chen, and Wei-Wei Tu. 2023. Tizero: Mastering multi-agent football with curriculum learning and self-play. *arXiv preprint arXiv:2302.07515*.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, and 1 others. 2024a. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024c. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.

Jianqiao Lu, Zhiyang Dou, Hongru Wang, Zeyu Cao, Jianbo Dai, Yunlong Feng, and Zhijiang Guo. 2024. Autopsv: Automated process-supervised verifier. *Advances in Neural Information Processing Systems*, 37:79935–79962.

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

Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jifeng Dai, Yu Qiao, and Xizhou Zhu. 2024. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. *arXiv preprint arXiv:2410.08202*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Feng Pan, Hanfei Zhang, Xuebao Li, Moyu Zhang, and Yang Ji. 2024. Achieving optimal trade-off for student dropout prediction with multi-objective reinforcement learning. *PeerJ Computer Science*, 10:e2034.

Anastasia Pentina, Viktoriia Sharmanska, and Christoph H Lampert. 2015. Curriculum learning of multiple tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5492–5500.

Kanghyun Ryu, Qiayuan Liao, Zhongyu Li, Payam Delgosha, Koushil Sreenath, and Negar Mehr. 2024. Curricullm: Automatic task curricula design for learning complex robot skills using large language models. *arXiv preprint arXiv:2409.18382*.

John Schulman and 1 others. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*.

Gaurav Srivastava, Shuxiang Cao, and Xuan Wang. 2025. Towards reasoning ability of small language models. *arXiv preprint arXiv:2502.11569*.

Mikihiro Tanaka, Takayuki Itamochi, Kenichi Narioka, Ikuro Sato, Yoshitaka Ushiku, and Tatsuya Harada. 2019. Generating easy-to-understand referring expressions for target identifications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5794–5803.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Mauricio Tec, Guojun Xiong, Haichuan Wang, Francesca Dominici, and Milind Tambe. 2025. Rule-bottleneck reinforcement learning: Joint explanation and decision optimization for resource allocation with language agents. *arXiv preprint arXiv:2502.10732*.

Carroll Wainwright and Ryan Lowe. 2023. Instructgpt: Training language models to follow instructions with human feedback. *GitHub repository*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Yu-Jie Wei, Hong-Peng Zhang, and Chang-Qiang Huang. 2023. Maneuver decision-making for autonomous air combat through curriculum learning and reinforcement learning with sparse rewards. *arXiv preprint arXiv:2302.05838*.

Zhiheng Xi, Wenxiang Chen, Boyang Hong, Senjie Jin, Rui Zheng, Wei He, Yiwen Ding, Shichun Liu, Xin Guo, Junzhe Wang, and 1 others. 2024. Training large language models for reasoning through reverse curriculum reinforcement learning. *arXiv preprint arXiv:2402.05808*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

Tong Yu, Yongcheng Jing, Xikun Zhang, Wentao Jiang, Wenjie Wu, Yingjie Wang, Wenbin Hu, Bo Du, and Dacheng Tao. 2025. Benchmarking reasoning robustness in large language models. *arXiv preprint arXiv:2503.04550*.

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

Tiantian Zhang, Kevin Zehua Shen, Zichuan Lin, Bo Yuan, Xueqian Wang, Xiu Li, and Deheng Ye. 2023. Replay-enhanced continual reinforcement learning. *arXiv preprint arXiv:2311.11557*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.