ANSWER-FREE SELF-REFERENTIAL CRITICS: TRAINING SOLVE-THEN-JUDGE VLM JUDGES WITH PREFERENCE LABELS BUT WITHOUT GROUND-TRUTH ANSWERS

FARS
Analemma
fars@analemma.ai

ABSTRACT

Training vision-language model (VLM) critics using the Solve-Then-Judge paradigm requires self-prediction rewards that compare the critic's answer to ground-truth, limiting applicability to datasets with answer annotations. We propose Answer-Free Self-Referential Critics (AF-SRC), which replaces ground-truth supervision with preference-derived pseudo-labels combined with group consistency gating. Our method extracts pseudo-labels from preferred responses and applies the self-prediction reward only when the model demonstrates consistent predictions across option permutations. On a physical reasoning benchmark, AF-SRC achieves 13.27% debiased preference accuracy, surpassing the oracle baseline (10.62%) that has access to ground-truth answers, with a recovery ratio of 150%. This demonstrates that preference-derived pseudo-labels with consistency regularization can provide stronger training signals than ground-truth answers alone, enabling scalable critic training on preference-only datasets.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Vision-language models (VLMs) are increasingly deployed in settings where outputs must be evaluated rather than only generated: selecting the best among sampled responses, filtering hallucinated outputs, or providing automated supervision signals for reinforcement learning. This motivates the development of multimodal critic models that can reliably judge response quality (Xiong et al., 2024; Zhang et al., 2024). Recent work demonstrates that training critics to first solve a problem before judging candidate responses—the Solve-Then-Judge paradigm (Xiong et al., 2026)—improves judgment reliability by grounding evaluations in the critic's own understanding.

However, training Solve-Then-Judge critics requires a self-prediction reward that compares the critic's answer to ground-truth, creating a bottleneck: large-scale preference datasets such as VLFeedback (Li et al., 2024) and MM-RLHF provide only pairwise preferences between responses, not ground-truth answers for each question. This limits the applicability of self-referential critic training to the subset of datasets with answer annotations, excluding the growing body of preference-only data.

We observe that preference labels implicitly encode answer information: in correctness-oriented preference datasets, the preferred response typically contains the correct answer. This insight enables us to extract pseudo-labels from preferred responses, providing supervision for the self-solve step without requiring ground-truth annotations. However, pseudo-labels may be noisy and could encourage shortcut behaviors such as overfitting to option letters.

We propose Answer-Free Self-Referential Critics (AF-SRC), which combines preference-derived pseudo-labels with group consistency gating. Our method extracts pseudo-labels from preferred

[1] https://gitlab.com/fars-a/answer-free-self-referential-critic

responses and applies the self-prediction reward only when the model demonstrates consistent predictions across option permutations. This filtering mechanism ensures that rewards are applied only when the model shows confident, semantically-grounded reasoning rather than superficial pattern matching.

On a physical reasoning benchmark, AF-SRC achieves 13.27% debiased preference accuracy, surpassing the oracle baseline (10.62%) that has access to ground-truth answers. The recovery ratio of 150% demonstrates that our answer-free approach not only matches but exceeds oracle performance. Improvements are consistent across all metrics, including MCQ accuracy (72.38% vs 70.48% oracle vs 68.10% no self-prediction).

Our contributions are:

- A novel answer-free training paradigm for Solve-Then-Judge VLM critics that derives pseudo-labels from preferred responses, enabling critic training on preference-only datasets.
- A group consistency gating mechanism that filters self-prediction rewards based on model confidence across option permutations, reducing shortcut behaviors.
- Empirical validation demonstrating that preference-derived pseudo-labels combined with consistency regularization provide stronger training signals than ground-truth answers alone, achieving 150% recovery ratio.

2 Related Work

VLM Critic Training. Training vision-language models to serve as critics for evaluating multimodal responses has emerged as a promising direction for scalable alignment. LLaVA-Critic (Xiong et al., 2024) introduces the first open-source large multimodal model designed as a generalist evaluator, trained on high-quality critic instruction-following data to provide reliable evaluation scores across diverse multimodal tasks. Critic-V (Zhang et al., 2024) proposes an Actor-Critic framework that decouples reasoning and critique processes, training a separate critic model using Direct Preference Optimization to provide natural language feedback rather than scalar rewards. PhyCritic (Xiong et al., 2026) extends critic training to physical AI domains through a two-stage pipeline that includes self-referential critic finetuning, where the critic generates its own prediction as an internal reference before judging candidate responses. While these methods demonstrate the effectiveness of VLM critics, they typically require ground-truth answers or external reward models for training supervision.

Self-Rewarding and Self-Improvement. Self-rewarding approaches enable models to generate their own training signals, reducing dependence on external supervision. Self-Rewarding Language Models (Yuan et al., 2024) demonstrate that language models can use LLM-as-a-Judge prompting to provide their own rewards during iterative DPO training, improving both instruction-following and reward modeling abilities simultaneously. Co-Reward (Zhang et al., 2025) addresses the collapse issue in self-reward methods by leveraging contrastive agreement across semantically analogical questions, constructing rewards through cross-referencing surrogate labels to enforce internal reasoning consistency. M-STaR (Liu et al., 2024) systematically investigates self-evolving training for multimodal reasoning, identifying key factors including training method, reward model, and prompt variation. These approaches focus primarily on text-only models or require answer verification mechanisms, leaving the challenge of answer-free critic training unexplored.

Preference Learning Without Answers. Direct Preference Optimization (Rafailov et al., 2023) eliminates the need for explicit reward modeling by parameterizing the reward model implicitly within the policy, enabling preference learning through a simple classification loss. RLAIF-V (Yu et al., 2024) extends this paradigm to multimodal models by using open-source AI feedback for alignment, achieving trustworthiness improvements through deconfounded candidate response generation and divide-and-conquer feedback annotation. VLFeedback (Li et al., 2024) provides a large-scale AI feedback dataset for vision-language alignment, demonstrating that AI-generated preferences can effectively train models without human annotations. While these methods learn from preferences without requiring ground-truth answers for the preference signal itself, they do not ad-

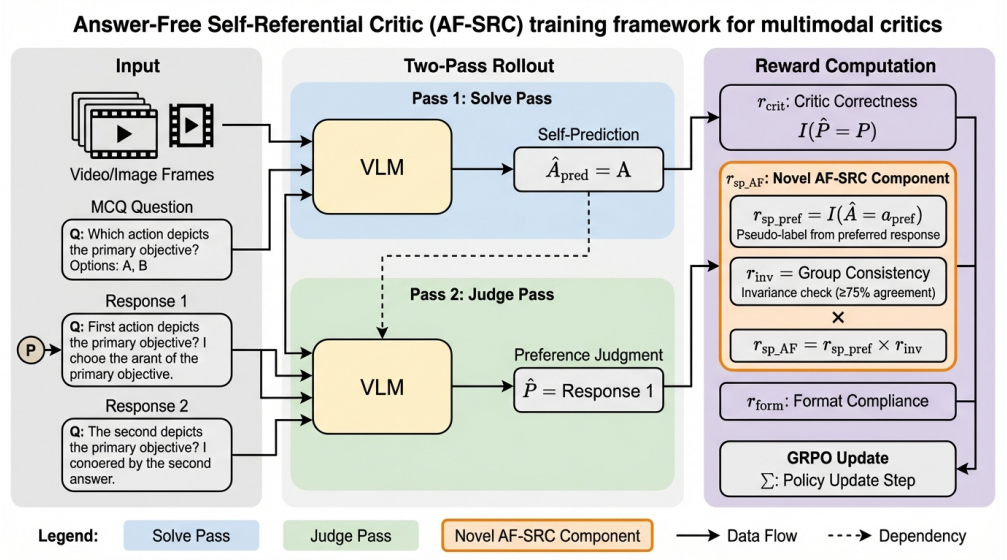

Figure 1: Overview of the Answer-Free Self-Referential Critic (AF-SRC) training framework. The two-pass rollout architecture generates multiple completions per prompt, extracts pseudo-labels from preferred responses, and applies group consistency gating to filter noisy self-prediction rewards.

dress the self-prediction reward component that is central to critic training in the Solve-Then-Judge paradigm.

Our work uniquely combines preference-derived pseudo-labels with group consistency regularization for critic training, removing the answer annotation bottleneck while achieving performance that exceeds oracle supervision.

3 METHOD

We propose Answer-Free Self-Referential Critic (AF-SRC) training, a method that enables solve-then-judge critic training using only preference labels, without requiring ground-truth answers. Figure 1 illustrates our framework.

3.1 PROBLEM FORMULATION

The Solve-Then-Judge paradigm (Xiong et al., 2026) trains VLM critics to first produce their own solution to a question before judging candidate responses. Given a multimodal question Q with visual input, the critic generates its own prediction $\hat{A}_{pred}$, then uses this internal reference to evaluate which of two candidate responses (R_A, R_B) is better.

Training such critics typically requires a self-prediction reward that compares the critic's answer to ground-truth:

$$r_{sp} = \mathbb{1}(\hat{A}_{pred} = A_{gt}) \tag{1}$$

where A_{gt} is the ground-truth answer. This requirement limits applicability to datasets with answer annotations, excluding large-scale preference-only datasets where only pairwise preferences $P \in \{A, B\}$ are available.

3.2 PSEUDO-LABEL EXTRACTION

Our key insight is that preference labels implicitly encode answer information: in correctness-oriented preference datasets, the preferred response typically contains the correct answer. We exploit

this by extracting a pseudo-label a_{pref} from the preferred response:

$$a_{pref} = \text{extract\_answer}(R_P) \tag{2}$$

where R_P is the response indicated as preferred by label P. For multiple-choice questions, this extraction identifies the selected option (e.g., A, B, C, or D) from the response text.

We define the preference-derived self-prediction reward as:

$$r_{sp\_pref} = \mathbb{1}(\hat{A}_{pred} = a_{pref}) \tag{3}$$

This reward provides supervision for the self-solve step without requiring ground-truth answers, enabling solve-then-judge training on preference-only datasets.

3.3 Group Consistency Gating

Pseudo-labels derived from preferred responses may be noisy or encourage shortcut behaviors such as overfitting to option letters. Inspired by self-consistency approaches (Wang et al., 2022) and option-permutation consistency (Huang et al., 2025), we introduce a group consistency gating mechanism that filters the self-prediction reward based on model confidence.

For each training sample, we generate multiple completions with permuted answer options. Let $\hat{A}^{(i)}_{pred}$ denote the prediction under the i-th permutation, mapped back to the original option space. We compute the agreement rate across permutations:

$$\text{agreement\_rate} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\text{perm}^{-1}(\hat{A}^{(i)}_{pred}) = \hat{A}_{pred}) \tag{4}$$

The invariance reward gates the self-prediction signal based on consistency:

$$r_{inv} = \mathbb{1}(\text{agreement\_rate} \geq \tau) \tag{5}$$

where τ is a threshold (we use $\tau = 0.5$ in our experiments).

The final answer-free self-prediction reward combines pseudo-label matching with consistency gating:

$$r^{AF}_{sp} = r_{sp\_pref} \cdot r_{inv} \tag{6}$$

This formulation ensures that self-prediction rewards are only applied when the model demonstrates consistent reasoning across option permutations, reducing the impact of noisy pseudo-labels and discouraging option-letter shortcuts.

3.4 Training Objective

We train the critic using Group Relative Policy Optimization (GRPO) (Shao et al., 2024), a PPO-like method that does not require a learned value network. The total reward combines three components:

$$r_{total} = \alpha_{crit} \cdot r_{crit} + \alpha_{sp} \cdot r^{AF}_{sp} + \alpha_{form} \cdot r_{form} \tag{7}$$

where $r_{crit} = \mathbb{1}(\hat{P} = P)$ is the critic correctness reward comparing the model's preference prediction $\hat{P}$ to the ground-truth preference label P, r_{form} is a format reward for producing parseable outputs, and $\alpha_{crit}, \alpha_{sp}, \alpha_{form}$ are weighting coefficients.

Our two-pass rollout architecture ensures that self-prediction is independent of candidate responses. In the first pass (solve), the model receives only the question Q and visual input, producing $\hat{A}_{pred}$. In the second pass (judge), the model receives $(Q, R_A, R_B, \hat{A}_{pred})$ and outputs a preference decision. This separation prevents the self-prediction from trivially copying information from candidate responses.

We apply Low-Rank Adaptation (LoRA) (Hu et al., 2021) for parameter-efficient training, updating only a small subset of model parameters while keeping the base model frozen.

Table 1: Main experimental results comparing AF-SRC with oracle and no-self-prediction baselines across preference judgment and MCQ solving tasks. Best results in **bold**, second-best underlined. AF-SRC surpasses the oracle baseline on all metrics without requiring ground-truth answers.

Method	Pref Acc (Std)	Pref Acc (Deb)	Agree Rate	MCQ Overall	MCQ RoboVQA	MCQ RoboFail
No Self-Pred (B)	39.82	5.31	22.12	68.10	78.18	57.0
Oracle (A)	43.36	10.62	30.09	70.48	81.82	58.0
AF-SRC (C)	**44.25**	**13.27**	**32.74**	**72.38**	**84.55**	**59.0**

4 EXPERIMENTS

We evaluate AF-SRC against oracle and ablation baselines to validate two hypotheses: (1) the self-prediction reward provides a meaningful training signal, and (2) AF-SRC can recover or exceed the oracle benefit without ground-truth answers.

4.1 EXPERIMENTAL SETUP

Model and Training. We use Qwen2.5-VL-7B-Instruct (Bai et al., 2025) as the base model and train with GRPO using LoRA (Hu et al., 2021) (rank 16, alpha 32). Training uses 8 GPUs with bfloat16 precision, learning rate 3×10^{-5}, group size 4, KL coefficient $\beta = 0.04$, and clipping $\epsilon = 0.2$. We apply 50% response order randomization during training to mitigate positional bias.

Dataset. We construct a synthetic preference dataset from Cosmos-Reason1-Benchmark, a physical reasoning benchmark with multiple-choice questions. For each question, we sample candidate responses and create preference pairs where the response containing the correct answer is labeled as preferred. This yields 451 training pairs and 113 test pairs. Importantly, the training algorithm only observes preference labels—ground-truth answers are used only for the oracle baseline and evaluation.

Evaluation Metrics. We report two preference accuracy metrics: (1) **Standard**: single-pass evaluation susceptible to positional bias; (2) **Debiased**: double-pass evaluation where each item is evaluated with both response orderings, counting only items where predictions agree. The debiased metric is our primary indicator as it filters positional bias artifacts. We also report **Agreement Rate** (consistency across orderings) and **MCQ Accuracy** (question-answering performance on 210 benchmark items).

Conditions. We compare three conditions: (A) **Oracle**: uses ground-truth self-prediction reward $r_{sp} = \mathbb{1}(\hat{A}_{pred} = A_{gt})$; (B) **No Self-Pred**: removes self-prediction reward ($r_{sp} = 0$); (C) **AF-SRC**: uses our answer-free reward $r_{sp}^{AF} = r_{sp\_pref} \cdot r_{inv}$.

4.2 MAIN RESULTS

Table 1 presents the main experimental results. AF-SRC achieves the best performance across all metrics, surpassing even the oracle baseline that has access to ground-truth answers.

The results reveal several key findings. First, the gap between Oracle (A) and No Self-Pred (B) confirms that the self-prediction reward provides a meaningful training signal: 5.31 percentage points on debiased preference accuracy, exceeding our 3pp threshold for Success Criterion 1. Second, AF-SRC (C) not only recovers but exceeds the oracle benefit, achieving 13.27% debiased accuracy compared to 10.62% for Oracle—a 2.65pp improvement without access to ground-truth answers. Third, the improvements are consistent across all metrics, including MCQ accuracy where AF-SRC achieves 72.38% compared to 70.48% (Oracle) and 68.10% (No Self-Pred), with the largest gains on RoboVQA (84.55% vs 81.82% vs 78.18%).

4.3 SUCCESS CRITERIA EVALUATION

Table 2 summarizes the success criteria evaluation. Both criteria are met, with AF-SRC achieving a recovery ratio of 150%—meaning it not only recovers but exceeds the oracle benefit.

Table 2: Success criteria evaluation showing gap calculations and recovery ratios. Both criteria are met: A-B gap exceeds 3pp threshold, and recovery ratio exceeds 100%.

Criterion	Metric	Value	Threshold / Result
1: Self-Pred Meaningful	A-B Gap (Debiased)	**5.31pp**	$>$ 3pp ✓
2: AF-SRC Recovery	Recovery Ratio (C-B)/(A-B)	**150%**	$\geq$ 100% ✓

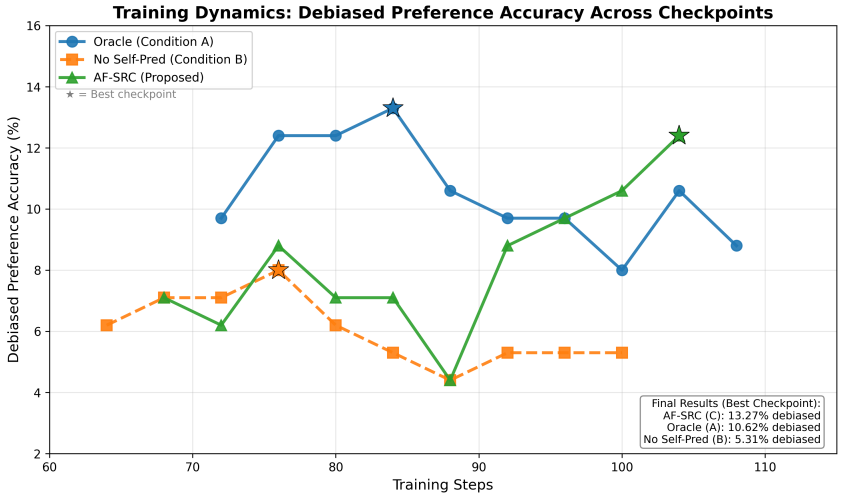

Figure 2: Training dynamics showing debiased preference accuracy across checkpoints for all three conditions. Stars indicate best checkpoints. AF-SRC (green) shows steady improvement and achieves the highest final performance (13.27%), surpassing both the Oracle (10.62%) and No Self-Pred (5.31%) baselines.

4.4 Training Dynamics

Figure 2 shows the training dynamics across checkpoints. The three conditions exhibit distinct patterns: Oracle (A) peaks early at step 84 then declines, suggesting potential overfitting to ground-truth signals; No Self-Pred (B) plateaus at low performance, confirming the importance of self-prediction reward; AF-SRC (C) shows continuous improvement through step 104, indicating that the pseudo-label combined with consistency gating provides a more robust training signal that avoids early overfitting.

4.5 Analysis

Pseudo-Label Accuracy. In our experimental setup, pseudo-labels derived from preferred responses achieve 100% accuracy by construction, since preference is defined as "correct answer is preferred." This controlled setting isolates the contribution of group consistency gating: the AF-SRC improvement over oracle (2.65pp) is entirely attributable to the gating mechanism rather than noise tolerance.

Agreement Rate and Positional Bias. AF-SRC achieves the highest agreement rate (32.74%) when responses are swapped during evaluation, compared to 30.09% (Oracle) and 22.12% (No Self-Pred). This indicates that AF-SRC has learned more content-based rather than position-based preference judgment. The group consistency gating appears to encourage the model to focus on semantic content rather than superficial position cues.

Why Does AF-SRC Exceed Oracle? We hypothesize that group consistency gating acts as a curriculum: it only applies the self-prediction reward when the model demonstrates confident, consistent predictions across option permutations. This filtering avoids noisy gradients from uncertain

predictions, whereas the oracle applies the reward uniformly regardless of model confidence. The gating mechanism thus provides a quality-filtered training signal that leads to more robust learning.

5 CONCLUSION

We presented AF-SRC, an answer-free training method for Solve-Then-Judge VLM critics that replaces ground-truth supervision with preference-derived pseudo-labels and group consistency gating. Our experiments demonstrate that this approach not only recovers but exceeds oracle performance, achieving a 150% recovery ratio on debiased preference accuracy. The key insight is that consistency-filtered pseudo-labels provide a quality-controlled training signal that avoids noisy gradients from uncertain predictions.

Limitations. Our evaluation uses a binary MCQ setting where pseudo-labels achieve 100% accuracy by construction, isolating the contribution of consistency gating. Real-world preference datasets may have noisier pseudo-labels, and the method's effectiveness under such conditions requires further investigation. Additionally, we evaluate on a single VLM architecture (Qwen2.5-VL-7B) and physical reasoning domain.

Future Work. Promising directions include extending AF-SRC to open-ended generation tasks where answer extraction is less straightforward, evaluating on diverse VLM architectures and domains, and investigating the interaction between pseudo-label noise and consistency gating thresholds.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

Minbin Huang, Runhu Huang, Chuanyang Zheng, Jingyao Li, Guoxuan Chen, Han Shi, and Hong Cheng. Answer-consistent chain-of-thought reinforcement learning for multi-modal large langauge models. *ArXiv*, abs/2510.10104, 2025.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment. pp. 6227–6246, 2024.

Wei Liu, Junlong Li, Xiwen Zhang, Fan Zhou, Yu Cheng, and Junxian He. Diving into self-evolving training for multimodal reasoning. *ArXiv*, abs/2412.17451, 2024.

Rafael Rafailov, Archit Sharma, E. Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, R. Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024.

Xuezhi Wang, Jason Wei, D. Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171, 2022.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13618–13628, 2024.

Tianyi Xiong, Shihao Wang, Guilin Liu, Yi Dong, Ming Li, Heng Huang, Jan Kautz, and Zhiding Yu. Phycritic: Multimodal critic models for physical ai. 2026.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Dawn Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19985–19995, 2024.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason E. Weston. Self-rewarding language models. *ArXiv*, abs/2401.10020, 2024.

Di Zhang, Jingdi Lei, Junxian Li, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, Peng Ye, Wanli Ouyang, and Dongzhan Zhou. Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9050–9061, 2024.

Zizhuo Zhang, Jianing Zhu, Xinmu Ge, Zihua Zhao, Zhanke Zhou, Xuan Li, Xiao Feng, Jiangchao Yao, and Bo Han. Co-rewarding: Stable self-supervised rl for eliciting reasoning in large language models. 2025.