IREASONER: Trajectory-Aware Intrinsic Reasoning Supervision for Self-Evolving Large Multimodal Models

Anonymous ACL submission

Abstract

Recent work shows that large multimodal models (LMMs) can self-improve from unlabeled data via self-play and intrinsic feedback. Yet existing self-evolving frameworks mainly reward final outcomes, leaving intermediate reasoning weakly constrained despite its importance for visually grounded decision making. We propose IREASONER, a self-evolving framework that improves an LMM's *implicit* reasoning by explicitly eliciting chain-of-thought (CoT) and rewarding its internal agreement. In a *Proposer–Solver* loop over unlabeled images, IREASONER augments outcome-level intrinsic rewards with a trajectory-aware signal defined over intermediate reasoning steps, providing learning signals that distinguish reasoning paths leading to the same answer without ground-truth labels or external judges. Starting from Qwen2.5-VL-7B, IREASONER yields up to +2.1 points across diverse multimodal reasoning benchmarks under fully unsupervised post-training. We hope this work serves as a starting point for reasoning-aware self-improvement in LMMs in purely unsupervised settings. Our code will be released.

1 Introduction

Self-improvement has become a practical way to push foundation models beyond supervised instruction tuning by generating training experiences and learning from internal feedback (Deng et al., 2025a; Huang et al., 2025; Srivastava et al., 2025; Ye et al., 2025; Fernando et al., 2024). In large multimodal models (LMMs), this idea has recently enabled self-evolving pipelines that train directly on streams of unlabeled images: a model proposes visually grounded questions, samples multiple solution attempts, and updates itself with intrinsic rewards computed from its own outputs (Thawakar et al., 2025; He et al., 2025). These approaches point to an appealing scaling path, since raw visual data is abundant while high-quality multimodal annotations and external reward systems remain costly to obtain (Dong et al., 2025; Lin et al., 2025).

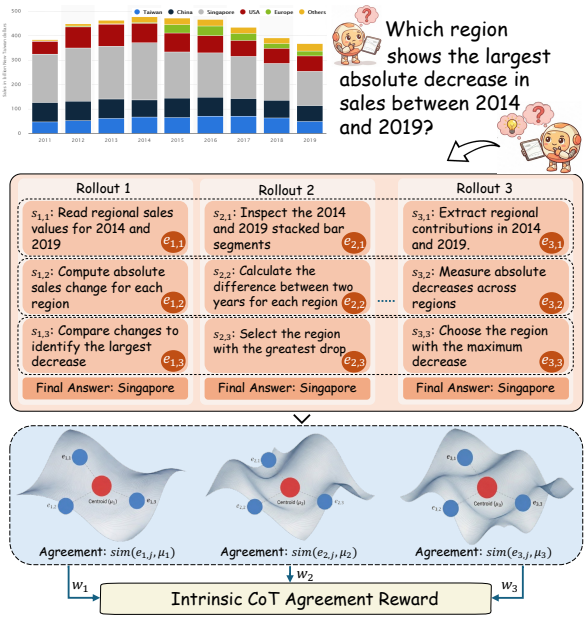

Figure 1: **IREASONER's intrinsic step-level CoT agreement.** Given an unlabeled image, the *Proposer* generates a visually grounded question, and the *Solver* samples N reasoning rollouts, each producing a CoT with multiple intermediate steps and a final answer (3 rollouts and 3 steps are shown here). Among rollouts in the dominant (majority-answer) group, we embed each step text $s_{i,j}$ into $e_{i,j}$ and form a per-step prototype μ_j. Step agreement is computed via similarity $\mathrm{sim}(e_{i,j}, \mu_j)$ and aggregated with higher weight on earlier, grounding-heavy steps ($w_1 > w_2 > w_3$) to produce a scalar Intrinsic CoT Agreement Reward.

A central limitation is that most self-improving LMM frameworks still define verification over final outcomes—such as a final answer, a whole response, a preference score, or a reconstruction score—while leaving intermediate reasoning weakly constrained (Deng et al., 2025b; Zhang et al., 2025a; Deng et al., 2025a). This matters even when the goal is not to build an explicit "reasoning model": instruction-tuned LMMs often rely on

implicit reasoning, and we can elicit step-by-step chains of thought (CoT) during training to shape that latent computation (Guo et al., 2025; Zhang et al., 2025b; Ge et al., 2023; Zhang et al., 2024d). Under outcome-only rewards, two responses that reach the same answer receive nearly the same learning signal, even if one is grounded in the image and the other relies on shortcuts or hallucinated intermediate claims that happen to cancel out. Outcome-only optimization therefore provides little direct pressure to stabilize the sequence of intermediate claims that drives reliable vision–language reasoning.

Prior work has made steady progress on self-improvement, but step-aware signals are still difficult to obtain in a purely unsupervised setting. Early approaches were developed mainly for large language models (LLMs) (Chen et al., 2025a; Zweiger et al., 2025; Lu et al., 2024) and later extended to LMMs using labeled supervision, external judges, or auxiliary reward systems (Zhou et al., 2024; Wang et al., 2025b; Li et al., 2025). Other approaches sidestep step supervision by using coarse consistency or reconstruction objectives to curate synthetic training data (Deng et al., 2025b; Zhang et al., 2025a). More recently, EvoLMM (Thawakar et al., 2025) and VisPlay (He et al., 2025) showed that purely unlabeled self-evolution from raw images is feasible; however, their intrinsic rewards still operate primarily at the answer/response level, leaving open the problem of how to evaluate and optimize CoT steps in multimodal self-evolution without labeled data or external supervision.

We address this gap by introducing IREASONER, a self-evolving framework that improves an LMM's *implicit* reasoning by eliciting explicit step-by-step CoT and rewarding its internal agreement. We follow the *Proposer–Solver* self-evolution backbone on unlabeled images (Thawakar et al., 2025; He et al., 2025), but differ in how the *Solver* is supervised: unlike prior methods whose intrinsic rewards are defined at the answer/response level, IREASONER introduces a trajectory-aware reward defined over intermediate CoT steps. As illustrated in Fig. 1, we group *Solver* rollouts by the dominant (majority) answer, align their CoT into step indices, and compute per-step prototypes from the dominant group. Each rollout is then rewarded by its step-wise similarity to these prototypes, with higher weight on early, grounding-heavy steps, yielding a single scalar Intrinsic CoT Agreement Reward. This reward is fully intrinsic and can drive policy-gradient updates without ground-truth labels, external judges, or verifiers, while distinguishing between reasoning trajectories that arrive at the same final answer but differ in intermediate claims.

To summarize, our contributions are as follows:

- We introduce IREASONER, a fully unsupervised self-evolving framework that brings intermediate reasoning into the optimization loop for *Proposer–Solver* self-evolution on unlabeled images.
- We propose an intrinsic CoT agreement reward that scores step-level alignment among *Solver* rollouts, providing trajectory-aware supervision that distinguishes reasoning paths leading to the same answer without labeled data or external judges.
- Starting from Qwen2.5-VL-7B, we empirically demonstrate that rewarding CoT improves self-evolving LMMs across diverse multimodal reasoning benchmarks, yielding gains of up to $+2.1$ points under fully unsupervised post-training.

2 Related Works

Self-Evolution in LMMs. In fully unsupervised image-only settings, EvoLMM (Thawakar et al., 2025) instantiates a cooperative *Proposer–Solver* trained with continuous self-consistency rewards, while VisPlay (He et al., 2025) alternates an image-conditioned questioner and a multimodal reasoner using group-relative rewards to balance difficulty and answer quality. Vision-Zero (Wang et al., 2025a) frames learning as strategic visual self-play over games generated from image pairs with RLVR-style updates. Other extensions broaden the loop: C2-Evo (Chen et al., 2025b) co-evolves synthetic multimodal data to calibrate training challenges, and Agent0-VL (Liu et al., 2025b) integrates tools for reasoning and self-verification. Vision-SR1 self-rewards a decomposed perception trace but still depends on ground-truth answers (Li et al., 2025). Complementary methods reduce hallucinations via preference-based signals, such as CSR's visually constrained rewards and SIMA's in-context visual self-critic (Zhou et al., 2024; Wang et al., 2025b). Separately, self-refinement pipelines filter synthetic IQA triplets via triangular consistency (Deng et al., 2025b) or bootstrap fine-grained perception through reconstruction and staged RL (Zhang et al., 2025a).

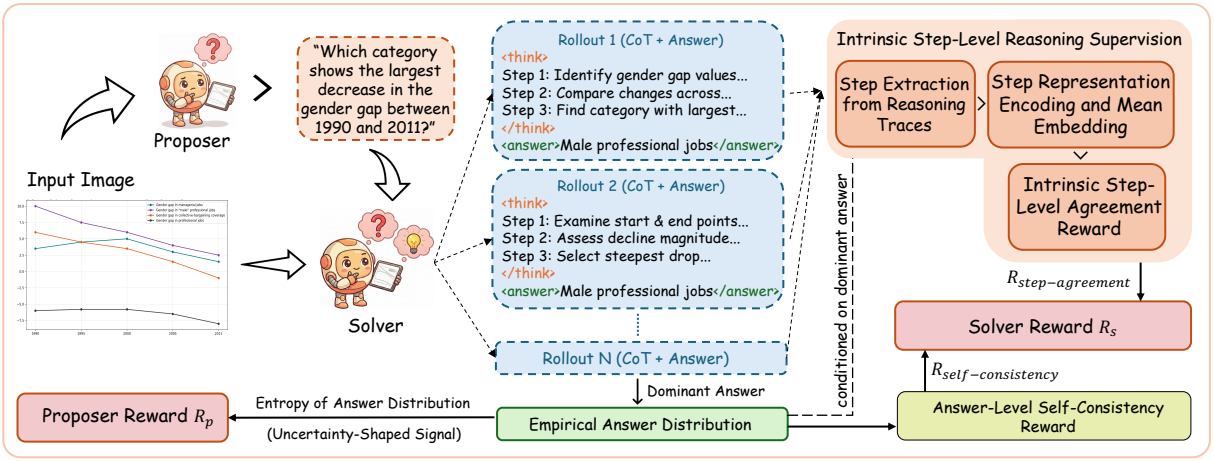

Figure 2: **Overview of our iReasoner pipeline.** From an unlabeled image x, a *Proposer* policy π_p generates a question q, and a *Solver* policy π_s produces N sampled reasoning rollouts, inducing an empirical answer distribution $p(a \mid x, q)$. The distribution entropy provides an uncertainty-shaped reward for the *Proposer* and selects the dominant answer group for *Solver*-side step supervision. The *Solver* reward combines answer-level self-consistency with an intrinsic step-level agreement signal computed over intermediate reasoning traces by (i) extracting numbered steps, (ii) embedding each step, and (iii) forming per-step prototypes within the dominant-answer group. This yields intrinsic step-level supervision without annotated question–answer pairs or external verifiers.

Optimizing Chain-of-Thought for Reasoning. A growing body of work explores how to elicit, control, and improve CoT reasoning in LLMs and LMMs. On the prompting and efficiency side, CAR (Lu et al., 2025) adaptively routes between short answers and long rationales using uncertainty signals, while token-control methods allocate reasoning budgets based on complexity to reduce overhead (Han et al., 2025). CoT quality can also be improved through context selection, by retrieving multimodal demonstrations via cross- and intra-modal similarity (Liu et al., 2025a). Other approaches focus on the reasoning trajectory itself. Multimodal-CoT (Zhang et al., 2024e) separates rationale generation from answer inference to improve convergence and reduce hallucinations, and R3V (Cheng et al., 2025) refines multimodal rationales through reflection and self-training. Training-time objectives align intermediate reasoning, including cascaded self-evaluation and filtering (Lv et al., 2025), preference optimization over CoT steps (Zhang et al., 2024c), and causal analyses that prune redundant steps or insert missing ones (Yu et al., 2025). Complementary studies examine when long CoTs emerge under SFT or RL and propose quantification frameworks such as reasoning boundaries, to guide CoT optimization (Yeo et al., 2025; Chen et al., 2024).

Despite this progress, the two lines above remain only loosely connected. Self-evolving LMM pipelines largely optimize outcome-level signals, while CoT optimization methods typically assume labeled data, external judges, or offline supervision to assess reasoning quality. As a result, intermediate reasoning cannot yet be evaluated or optimized in a purely unsupervised self-evolution loop; iReasoner bridges this gap by turning cross-rollout step agreement into an intrinsic reward that directly trains the reasoning process during self-improvement.

3 Method

3.1 Overview

We introduce iReasoner, a fully unsupervised self-evolving framework that improves an LMM by explicitly optimizing the *Solver*'s CoT, not only its final answers. Given unlabeled images $x \sim \mathcal{D}$, we adopt the *Proposer–Solver* self-evolution regime from prior work: a *Proposer* π_p generates a visually grounded question $q \sim \pi_p(\cdot \mid x)$ and a *Solver* π_s samples N rollouts $y_i \sim \pi_s(\cdot \mid x, q)$ (Fig. 2). Each *Solver* rollout is structured as $y_i = \langle$<think> t_i </think><answer> a_i </answer>$\rangle$, where t_i is an explicit multi-step trace and a_i is the extracted answer. Sampling induces an empirical answer distribution over normalized answers,

$$p(a \mid x, q) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[a_i = a], \qquad (1)$$

which provides intrinsic supervision without labeled question–answer pairs or external verifiers.

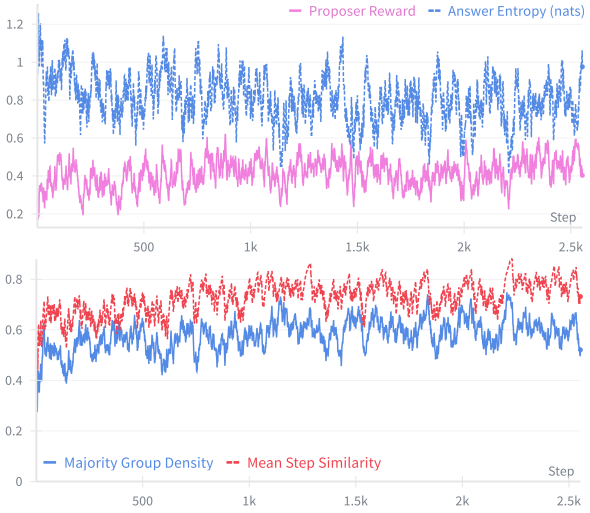

Figure 3: **Question shaping and *Solver* reasoning behavior over training steps.** *(Top) Proposer* reward remains stable (around 0.3–0.5) while answer entropy stays in a moderate band (roughly 0.6–1.1 nats), consistent with sustained intermediate difficulty rather than degeneracy. *(Bottom)* Majority-group density and mean step similarity increase over training, indicating a larger fraction of *Solver* samples agree on the dominant answer and that their intermediate steps become more aligned.

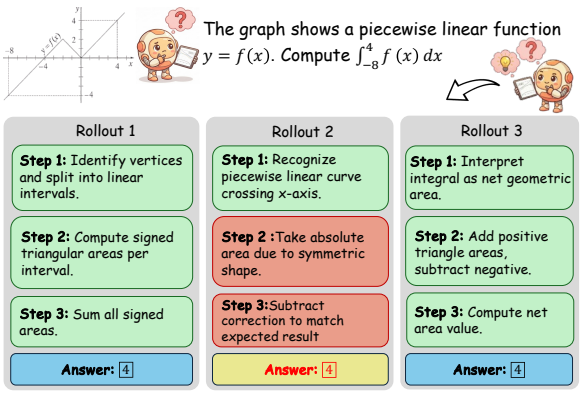

Figure 4: **Outcome-only self-consistency treats distinct CoTs similarly.** For the same image-question pair, three *Solver* rollouts produce the same final answer, but their intermediate steps differ: Rollouts 1 and 3 follow a consistent signed-area decomposition, while Rollout 2 deviates via incorrect intermediate claims yet still ends at the same answer. Since outcome-only intrinsic rewards depend only on answer agreement, these rollouts receive nearly identical learning signal despite qualitatively different reasoning traces, motivating step-aware supervision in IREASONER.

As in EvoLMM (Thawakar et al., 2025), outcome-level intrinsic verification is derived from $p(a \mid x, q)$: the *Solver* is rewarded for answer self-consistency, while the *Proposer* is shaped toward intermediate-difficulty questions using the answer entropy $H(x, q) = -\sum_a p(a \mid x, q) \log p(a \mid x, q)$. We adopt this *Proposer*-side uncertainty shaping and observe stable *Proposer* reward and sustained non-degenerate answer entropy throughout training (Fig. 3, top), indicating that the self-evolution loop remains in a meaningful difficulty regime rather than collapsing to trivial questions.

However, outcome-only supervision leaves intermediate reasoning weakly constrained: rollouts that share an answer can receive nearly identical learning signal despite very different CoTs (as illustrated in Fig. 4). To this end, IREASONER adds trajectory-aware supervision by an intrinsic mechanism that scores agreement of intermediate steps among rollouts that converge to the same answer, and integrating it into *Solver* optimization (Fig. 2).

3.2 Intrinsic CoT Agreement Reward

Our goal is to make intermediate reasoning learnable under unsupervised self-evolution. For a fixed (x, q), strong solutions should not only agree on the final answer, but also reuse consistent intermediate claims—especially in early grounding steps. We therefore measure step-wise agreement within rollouts that already agree on the answer.

We view each *Solver* trace as steps $t_i = (s_{i,1}, \ldots, s_{i,J_i})$, where $s_{i,j}$ is the j-th intermediate claim. To make agreement meaningful, we elicit a consistent step interface so that the same index j plays a comparable role across rollouts (e.g., early grounding vs. later computation). Let $\hat{a} = \arg\max_a p(a \mid x, q)$ be the dominant answer under Eq. (1), and let

$$\mathcal{G} = \{i \mid a_i = \hat{a}\}, \qquad \rho = \left(\frac{|\mathcal{G}|}{N}\right)^{\gamma}, \ \gamma \geq 0 \quad (2)$$

denote the dominant-answer group and a density-based reliability factor. Conditioning on $\mathcal{G}$ anchors step agreement to a stable outcome and avoids reinforcing agreement among off-target rollouts; ρ downweights step supervision when few samples agree.

To compare steps, we embed each step into $e_{i,j} = f(s_{i,j}) \in \mathbb{R}^d$ using the model's internal text representations. In our implementation, $f(\cdot)$ is the ℓ_2-normalized mean of input-token embeddings for the step text (with a fixed token budget), which provides a lightweight semantic representation aligned with the *Solver*'s language space. For

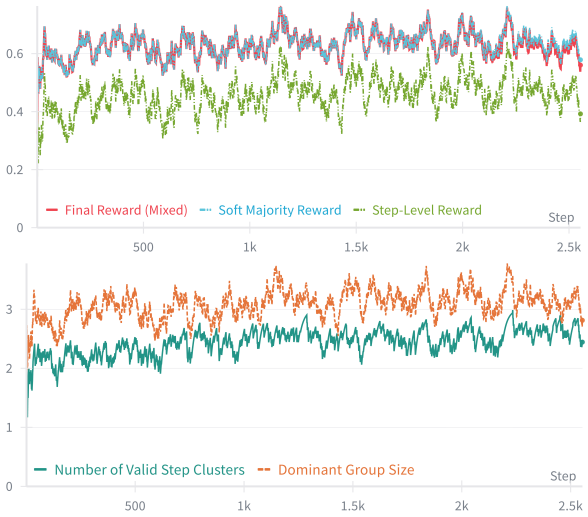

Figure 5: **Reward components and step-structure trends over training steps.** *(Top)* The final mixed *Solver* reward stays in the high range (about 0.6–0.7) while the step-level term provides additional signal beyond soft-majority answer consistency. *(Bottom)* The dominant-answer group size and the number of valid step positions used for agreement increase into the 2–3 range, indicating that an increasing portion of the trace is consistently comparable across samples.

each index j, we form a dominant-group prototype

$$\mu_j = \frac{1}{|\mathcal{G}_j|} \sum_{i \in \mathcal{G}_j} e_{i,j}, \qquad \mathcal{G}_j = \{i \in \mathcal{G} \mid j \le J_i\}, \tag{3}$$

and score per-step agreement by cosine similarity $r_{i,j} = \mathrm{sim}(e_{i,j}, \mu_j) \in [-1, 1]$. Because early steps typically carry grounding and setup, we aggregate with decaying weights $w_1 > w_2 > \cdots$:

$$\tilde{r}_i^{\text{step}} = \sum_{j \le J_i} w_j \, r_{i,j}. \tag{4}$$

Finally, we define the Intrinsic CoT Agreement Reward as $r_i^{\text{step}} = \rho \cdot \tilde{r}_i^{\text{step}}$. Both majority-group density and mean step similarity increase over training (Fig. 3, bottom), consistent with progressively stabilized *Solver* reasoning traces.

3.3 Step-Aware Self-Evolution via Reward Integration

We train the *Solver* by augmenting outcome-level self-consistency with intrinsic CoT agreement. Given Eq. (1), each rollout receives an answer-level reward

$$r_i^{\text{ans}} = p(a_i \mid x, q)^{\alpha} \left(1 - \eta \, \bar{\ell}_i\right), \tag{5}$$

where $\alpha > 0$ controls sharpness, $\bar{\ell}_i \in [0, 1]$ is a normalized excess pre-answer length relative to a target budget, and $\eta \ge 0$ sets the length penalty strength. IREASONER combines Eq. (5) with the step reward:

$$r_i^{\text{sol}} = (1 - \lambda(t))\, r_i^{\text{ans}} + \lambda(t)\, r_i^{\text{step}}, \tag{6}$$

where $\lambda(t) \in [0, 1]$ increases over training with a warmup and ramp. Early updates rely more on answer-level self-consistency, while later updates place more weight on step agreement once the dominant group and step positions are more reliably populated (Fig. 5, bottom). This is also reflected in the reward decomposition (Fig. 5, top), where the mixed reward remains high while the step-level term provides complementary signal beyond answer-level agreement.

We train both the *Solver* and *Proposer* with KL-regularized policy gradients against a frozen reference policy, which stabilizes self-evolution by limiting drift while preserving reward-driven updates. For the *Solver*, we apply REINFORCE (Williams, 1992) with an EMA baseline b_s:

$$\mathcal{L}_s(\theta) = -\mathbb{E}_{y \sim \pi_s(\cdot \mid x,q)} \left[(r^{\text{sol}}(y) - b_s) \, \log \pi_s(y \mid x, q) \right] + \beta_s \, \mathbb{E}[\mathrm{KL}(\pi_s(\cdot \mid x, q) \,\|\, \pi_{\text{ref}}(\cdot \mid x, q))], \tag{7}$$

The *Proposer* is optimized with the same objective form (baseline b_p and coefficient β_p), using the entropy-shaped reward $r^p(x, q) = g(H(x, q))$. To maintain a target KL budget, we adapt each coefficient online via a multiplicative controller:

$$\beta \leftarrow \mathrm{clip}\left(\beta \cdot \exp\left(\eta \, \frac{\mathrm{KL} - \tau}{\tau} \right), \, \beta_{\min}, \, \beta_{\max} \right), \tag{8}$$

which strengthens regularization when the observed KL exceeds τ and relaxes it otherwise.

4 Experiments

4.1 Experimental Setup

Training details. Our training pool contains 2.5k images sampled from six widely used sources (ChartQA (Masry et al., 2022), AI2D (Kembhavi et al., 2016), InfoGraphic-VQA (Mathew et al., 2022), PlotQA (Methani et al., 2020), ChartX (Xia et al., 2025), Geometry3K (Lu et al., 2021)). We train on images only; no question–answer pairs, captions, metadata, or external reward models are used at any stage.

We initialize from Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and train in the *Proposer–Solver* self-evolution regime using lightweight LoRA (Hu

Model	General Visual Understanding				Visual Math			
	InfoGraphic-VQA$_{val}$	AI2D	ScienceQA	MMMU$_{val}$	ChartQA	MathVista	MathVision	MathVerse
Vision-Zero† (Wang et al., 2025a)	80.35	82.64	88.50	51.44	84.24	68.43	23.96	43.86
VisPlay* (He et al., 2025)	–	–	–	38.27	–	–	**31.15**	39.14
Qwen2.5-VL-7B (Baseline) (Bai et al., 2025)	80.44	82.61	88.30	51.11	84.00	68.47	23.91	43.78
Qwen2.5-VL-7B w/ Discrete Reward (Thawakar et al., 2025)	80.52	82.18	87.98	50.84	84.62	68.88	22.52	42.10
EvoLMM (Thawakar et al., 2025)	**81.06**	**83.41**	**89.50**	**52.01**	**86.70**	**70.52**	24.81	**44.88**
Qwen2.5-VL-7B w/ Discrete Reward + Step-level Majority	80.78	82.95	88.92	51.48	85.42	69.31	24.12	44.18
Qwen2.5-VL-7B w/ Cont. Reward + Step-level Majority (Ours)	**81.56**	**83.89**	**89.92**	**52.37**	**85.78**	**69.74**	**25.29**	**45.91**

Table 1: **Evaluation results across eight multimodal reasoning benchmarks.** Benchmarks are grouped into general visual understanding and visual mathematics tasks. Best and second-best results are highlighted. Methods marked with (†) use external supervision. Methods marked with (*) report results using LLM-as-a-judge evaluation.

Ablation	General Visual Understanding				Visual Math			
	InfoGraphic-VQA$_{val}$	AI2D	ScienceQA	MMMU$_{val}$	ChartQA	MathVista	MathVision	MathVerse
Qwen2.5-VL-7B (Baseline)	80.44	82.61	88.30	51.11	84.00	68.47	23.91	43.78
Step-level majority (Ours, Full)	**81.56**	**83.89**	**89.92**	**52.37**	**85.78**	**69.74**	**25.29**	**45.91**
Soft majority reward only (EvoLMM) (Thawakar et al., 2025)	81.12	83.36	89.41	51.92	86.64	70.41	24.62	44.71
Step-level reward only	80.61	82.69	88.44	50.98	84.38	68.73	24.18	43.87
Step-Level Mechanism Design								
w/o Warmup schedule	81.04	83.21	89.26	51.74	85.02	68.97	24.63	45.11
w/o Position decay	81.29	83.58	89.55	52.02	85.41	69.34	25.02	45.49
w/o Density weighting	81.18	83.46	89.47	51.88	85.29	69.19	24.91	45.32
Reward Shaping Components								
w/o Length penalty	81.37	83.66	89.63	52.11	85.58	69.49	25.11	45.61
Soft majority $\gamma = 0.5$	81.09	83.31	89.34	51.71	85.21	69.07	24.72	45.12
Soft majority $\gamma = 1.0$	80.83	83.08	89.02	51.49	84.96	68.82	24.51	44.88

Table 2: **Ablation study of intrinsic reasoning supervision across eight benchmarks.** The full configuration achieves the strongest overall performance across both task groups. Answer stability alone performs best on highly verifiable benchmarks, while adding step-wise reasoning reward improves transfer on tasks where intermediate structure is informative. Removing mechanism or shaping components leads to consistent regressions, with warmup having the largest effect.

et al., 2022) adapters for both roles while keeping the backbone frozen. For each image, the *Proposer* samples one question and the *Solver* samples N=5 reasoning rollouts. The *Proposer* is updated every 5 iterations. The *Solver* is optimized with KL-regularized REINFORCE using our mixed intrinsic reward (answer stability + step-level agreement), with a warmup schedule that ramps the step component from 0 to a maximum weight of 0.7. We train for 2.5k steps using AdamW (learning rate 10^{-6}, weight decay 0.01, gradient clipping 1.0) in bfloat16 on 8× AMD MI250X GPUs; the full run completes in approximately 35 hours.

Evaluation protocol. We evaluate IREASONER trained from an instruction-tuned seed model (Qwen2.5-VL-7B (Bai et al., 2025)) using an intrinsic CoT–guided RL pipeline in a fully unlabeled setting. We evaluate on eight multimodal reasoning benchmarks: ChartQA (Masry et al., 2022), MathVista (Lu et al., 2023), MathVision (Wang et al., 2024), MathVerse (Zhang et al., 2024b), InfoGraphic-VQA (Mathew et al., 2022), AI2D (Kembhavi et al., 2016), ScienceQA (Lu et al., 2022), and MMMU (Yue et al., 2024). We use the official evaluation splits and each benchmark's standard accuracy metric. All models are evaluated with identical inference settings (no task-specific tuning), so differences reflect self-evolution rather than evaluation-time customization. Evaluations are run with lmms-eval (Zhang et al., 2024a) on AMD MI250X GPUs using bfloat16 for consistency with training.

4.2 Main Benchmark Results

Table 1 reports results on eight multimodal reasoning benchmarks spanning general visual understanding and visual mathematics. We make three primary observations.

IREASONER consistently improves over the seed model across both task groups. Across general visual understanding benchmarks, IREASONER improves over Qwen2.5-VL-7B by +1.12 on InfoGraphic-VQA, +1.28 on AI2D, +1.62 on ScienceQA, and +1.26 on MMMU, corresponding to an average gain of +1.32. On visual mathematics benchmarks, IREASONER improves on all four datasets (+1.64 on average), with the largest gain on MathVerse (+2.13) and a consistent improvement on MathVision (+1.38).

Step-wise reasoning reward improves general-purpose transfer beyond answer-level agreement. Compared to EvoLMM (Thawakar et al.,

	General Visual Understanding				Visual Math			
Max Reasoning Steps	**InfoGraphic-VQA$_{val}$**	**AI2D**	**ScienceQA**	**MMMU$_{val}$**	**ChartQA**	**MathVista**	**MathVision**	**MathVerse**
4 steps	80.92	83.12	89.21	51.82	85.10	69.08	24.71	45.02
6 steps	81.31	83.61	89.68	52.21	85.54	69.52	25.06	45.62
8 steps (default)	**81.56**	**83.89**	**89.92**	**52.37**	**85.78**	**69.74**	**25.29**	**45.91**
10 steps	81.42	83.74	89.79	52.26	85.63	69.61	25.12	45.71

Table 3: **Sensitivity to the maximum number of reasoning steps.** We report performance across all eight benchmarks while varying the maximum number of parsed reasoning steps used to extract intermediate reasoning structure. Smaller step budgets truncate useful intermediate information, while overly large budgets can introduce noisy or redundant steps. The default setting of 8 steps used in our main experiments provides a strong balance and achieves robust performance across both general visual understanding and visual mathematics tasks.

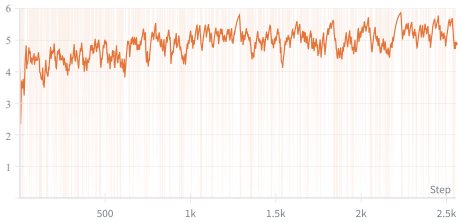

Figure 6: **Evolution of reasoning step usage over training.** Average number of extracted reasoning steps per solution (capped at eight), which gradually stabilizes over training, reflecting more consistent and structured intermediate reasoning.

2025), IREASONER improves all four general benchmarks: InfoGraphic-VQA (+0.50), AI2D (+0.48), ScienceQA (+0.42), and MMMU (+0.36). This pattern is strongest on tasks where multiple intermediate reasoning traces can reach the same final answer (such as those in MMMU and MathVerse), showing that cross-trajectory step agreement provides supervision that is not captured by answer-level consistency alone. On visual math benchmarks, IREASONER improves MathVision (24.81→25.29) and MathVerse (44.88→45.91), while EvoLMM remains stronger on ChartQA and MathVista. This contrast is consistent with answer-stability rewards being particularly effective on highly verifiable short-answer settings, whereas step-level alignment contributes more when intermediate structure is informative and transferable.

Step-level agreement requires continuous rewards to be effective. To isolate the interaction between reward type and step-level supervision, we include a discrete-reward variant augmented with step-level agreement. While adding step-level agreement to a discrete reward improves over the discrete-only baseline (e.g., InfoGraphic-VQA 80.52→80.78; MathVerse 42.10→44.18), the gains remain limited and do not match those achieved by IREASONER. In contrast, combining step-level agreement with continuous intrinsic rewards yields consistent improvements across all eight benchmarks, indicating that step-level signals benefit from smoother credit assignment, where partial alignment in intermediate steps can be rewarded even when answers are not perfectly consistent.

Comparison to prior self-evolving LMMs. We also compare against other self-evolving LMMs under their reported evaluation protocols. VisPlay (He et al., 2025), which operates in a similar unsupervised self-evolving regime, achieves higher performance on MathVision, while IREASONER substantially outperforms VisPlay on MMMU$_{val}$ (52.37 vs. 38.27) and MathVerse (45.91 vs. 39.14). Vision-Zero (Wang et al., 2025a) is included for context but relies on external supervision and is therefore not directly comparable.

4.3 Ablation Study

Table 2 ablates the two intrinsic rewards and key design choices in cross-trajectory step alignment.

Answer stability is strongest on highly verifiable benchmarks, while step-wise reward improves transfer. Using only the answer-stability reward yields large gains on ChartQA and MathVista, improving over the seed model by +2.64 (84.00→86.64) and +1.94 (68.47→70.41). Adding step-wise reward improves benchmarks where intermediate structure matters: relative to answer stability alone, IREASONER improves InfoGraphic-VQA by +0.44, AI2D by +0.53, ScienceQA by +0.51, MMMU by +0.45, and also increases MathVision by +0.67 and MathVerse by +1.20. At the same time, it is slightly lower on ChartQA (-0.86) and MathVista (-0.67), consistent with a trade-off between answer-level agreement and enforcing step-level structure on highly verifiable tasks.

Step-wise reward alone is weak, but becomes effective with answer stability. Using only step-wise reward yields small gains over the seed model (e.g., +0.17 on InfoGraphic-VQA; +0.09 on MathVerse) and does not improve MMMU. In contrast, combining both signals substantially improves over step-wise reward alone (e.g., +1.48 on ScienceQA; +2.04 on MathVerse). Overall, cross-trajectory step alignment benefits from an answer-level signal that reduces reward noise early in training.

Mechanism and shaping choices matter, and

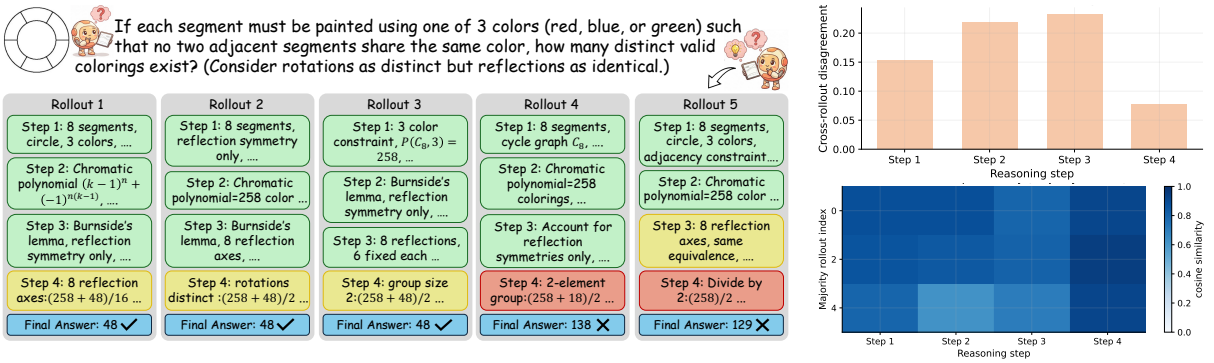

Figure 7: **Majority-group rollouts and step-level agreement diagnostics.** Left: qualitative sample of an image, *Proposer*-generated question and five *Solver* rollouts with four reasoning steps. Right (top): per-step disagreement D_j computed from leave-one-out similarities. Right (bottom): leave-one-out step similarity heatmap $S_{i,j}$ over rollouts $i \in \mathcal{G}$ and step indices j (higher indicates greater consistency with the group).

warmup is most important. Removing the warmup schedule produces the largest and most uniform regression (average -0.68 across benchmarks), including -0.76 on ChartQA and -0.80 on MathVerse. Removing position decay or density weighting also reduces performance with smaller but consistent drops. Reward shaping has systematic effects: removing the length penalty reduces MathVerse by -0.30, while removing shaping entirely (γ=1.0) leads to larger degradations (e.g., -0.90 on ScienceQA and -1.03 on MathVerse).

4.4 Training Dynamics and Reasoning-Trace Analysis

Sensitivity to the step budget. We analyze sensitivity to the maximum number of intermediate reasoning steps used for cross-trajectory alignment. Table 3 shows that increasing the step budget from 4 to 8 yields consistent gains (e.g., MathVerse 45.02→45.91; AI2D 83.12→83.89), while increasing beyond 8 gives diminishing returns and can slightly regress, consistent with longer traces adding redundant or noisier steps for alignment. Figure 6 further shows that the average number of extracted steps per rollout (capped at eight) stabilizes over training, suggesting that intrinsic reasoning supervision promotes more consistent trace structure rather than unbounded growth.

Within-mode divergence under answer agreement. Even when answers agree, intermediate reasoning can vary substantially, as rollouts in the dominant-answer group $\mathcal{G}$ may share the same final answer while diverging on key steps (Fig. 7 (left)). To localize where this occurs, we measure step agreement within $\mathcal{G}$ using the same step parsing and lightweight text embeddings as our step-level reward. For each rollout $i \in \mathcal{G}$ and step index j, we compute a leave-one-out similarity $S_{i,j} = \cos(e_{i,j}, \mu_{-i,j})$, where $\mu_{-i,j}$ is the mean step-j embedding over the other majority rollouts. Fig. 7 (bottom right) visualizes $S_{i,j}$ and shows that even under answer agreement, some rollouts deviate sharply at specific step indices while remaining aligned elsewhere. Aggregating across rollouts gives a depth-wise disagreement profile $D_j = 1 - \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} S_{i,j}$. As shown in Fig. 7 (top right), disagreement concentrates in the middle of the trace (steps 2–3) and is lower in later steps, supporting that outcome-only rewards cannot distinguish stable from unstable reasoning within the dominant answer mode, whereas step-level objectives directly target trace parts where rollouts drift.

5 Conclusion

We presented IREASONER, a self-evolving post-training framework for LMMs that brings intermediate reasoning into the optimization loop in a fully unlabeled setting. IREASONER introduces a trajectory-aware intrinsic signal that aligns *Solver* traces across responses that converge to the same answer, addressing a key limitation of outcome-only self-consistency where identical answers can arise from unstable step sequences. Starting from Qwen2.5-VL-7B and training exclusively on unlabeled images, IREASONER yields consistent gains across diverse multimodal reasoning benchmarks and shows that step-aware intrinsic supervision can improve transfer beyond answer-level agreement alone. We hope this work serves as a useful starting point for reasoning-aware self-improvement and more open-ended self-generated curricula under minimal supervision in multimodal systems.

Limitations

IREASONER uses only intrinsic signals derived from the model's own samples. As a result, it cannot directly optimize for external correctness: when the dominant-answer group is confidently wrong (e.g., due to early-training noise or perception failures), step-level agreement may reinforce internally consistent but incorrect reasoning. Mitigating this failure mode without introducing external judges remains open. Our work is also limited in scale and coverage. We train for 2.5k self-evolution steps on 2.5k unlabeled images and report results from a single backbone. Longer runs, larger and more diverse unlabeled image streams, and additional model families are needed to better characterize stability and scaling. Finally, our training procedure assumes access to model internals to compute log-probability objectives and KL regularization against a reference policy. This makes the approach most applicable to open-weight models and less directly transferable to black-box systems that do not expose token-level likelihoods.

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, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Lili Chen, Mihir Prabhudesai, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025a. Self-questioning language models. *Preprint*, arXiv:2508.03682.

Qiguang Chen, Libo Qin, Jiaqi Wang, Jinxuan Zhou, and Wanxiang Che. 2024. Unlocking the capabilities of thought: A reasoning boundary framework to quantify and optimize chain-of-thought. *Preprint*, arXiv:2410.05695.

Xiuwei Chen, Wentao Hu, Hanhui Li, Jun Zhou, Zisheng Chen, Meng Cao, Yihan Zeng, Kui Zhang, Yu-Jie Yuan, Jianhua Han, Hang Xu, and Xiaodan Liang. 2025b. C2-evo: Co-evolving multimodal data and model for self-improving reasoning. *Preprint*, arXiv:2507.16518.

Kanzhi Cheng, Li YanTao, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. 2025. Vision-language models can self-improve reasoning via reflection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8876–8892, Albuquerque, New Mexico. Association for Computational Linguistics.

Shijian Deng, Kai Wang, Tianyu Yang, Harsh Singh, and Yapeng Tian. 2025a. Self-improvement in multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1987–2006, Suzhou, China. Association for Computational Linguistics.

Yunlong Deng, Guangyi Chen, Tianpei Gu, Lingjing Kong, Yan Li, Zeyu Tang, and Kun Zhang. 2025b. Towards self-refinement of vision-language models with triangular consistency. *Preprint*, arXiv:2510.10487.

Hao Dong, Lijun Sheng, Jian Liang, Ran He, Eleni Chatzi, and Olga Fink. 2025. Adapting vision-language models without labels: A comprehensive survey. *Preprint*, arXiv:2508.05547.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2024. Promptbreeder: self-referential self-improvement via prompt evolution. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Jiaxin Ge, Hongyin Luo, Siyuan Qian, Yulu Gan, Jie Fu, and Shanghang Zhang. 2023. Chain of thought prompt tuning in vision language models. *Preprint*, arXiv:2304.07919.

Jiawei Guo, Tianyu Zheng, Yizhi Li, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Graham Neubig, Wenhu Chen, and Xiang Yue. 2025. MAmmoTH-VL: Eliciting multimodal reasoning with instruction tuning at scale. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13869–13920, Vienna, Austria. Association for Computational Linguistics.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. 2025. Token-budget-aware llm reasoning. *Preprint*, arXiv:2412.18547.

Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. 2025. Visplay: Self-evolving vision-language models from images. *Preprint*, arXiv:2511.15661.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Xu Huang, Weiwen Liu, Xingshan Zeng, Yuefeng Huang, Xinlong Hao, Yuxian Wang, Yirong Zeng, Chuhan Wu, Yasheng Wang, Ruiming Tang, and Defu Lian. 2025. Toolace-dev: Self-improving tool learning via decomposition and evolution. *Preprint*, arXiv:2505.07512.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.

Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, and Dong Yu. 2025. Self-rewarding vision-language model via reasoning decomposition. *Preprint*, arXiv:2508.19652.

Ci-Siang Lin, Min-Hung Chen, Yu-Yang Sheng, and Yu-Chiang Frank Wang. 2025. Leaml: Label-efficient adaptation to out-of-distribution visual tasks for multimodal large language models. *Preprint*, arXiv:2510.03232.

Bingshuai Liu, Chenyang Lyu, Zijun Min, Zhanyu Wang, Jinsong Su, and Longyue Wang. 2025a. Retrieval-augmented multi-modal chain-of-thoughts reasoning for large language models. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Jiaqi Liu, Kaiwen Xiong, Peng Xia, Yiyang Zhou, Haonian Ji, Lu Feng, Siwei Han, Mingyu Ding, and Huaxiu Yao. 2025b. Agent0-vl: Exploring self-evolving agent for tool-integrated vision-language reasoning. *Preprint*, arXiv:2511.19900.

Jianqiao Lu, Wanjun Zhong, Wenyong Huang, Yufei Wang, Qi Zhu, Fei Mi, Baojun Wang, Weichao Wang, Xingshan Zeng, Lifeng Shang, Xin Jiang, and Qun Liu. 2024. Self: Self-evolution with language feedback. *Preprint*, arXiv:2310.00533.

Jinghui Lu, Haiyang Yu, Siliang Xu, Shiwei Ran, Guozhi Tang, Siqi Wang, Bin Shan, Teng Fu, Hao Feng, Jingqun Tang, Han Wang, and Can Huang. 2025. Prolonged reasoning is not all you need: Certainty-based adaptive routing for efficient llm/mllm reasoning. *Preprint*, arXiv:2505.15154.

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*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Preprint*, arXiv:2209.09513.

Zheqi Lv, Wenkai Wang, Jiawei Wang, Shengyu Zhang, and Fei Wu. 2025. Cascaded self-evaluation augmented training for lightweight multimodal llms. *Preprint*, arXiv:2501.05662.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536.

Gaurav Srivastava, Zhenyu Bi, Meng Lu, and Xuan Wang. 2025. DEBATE, TRAIN, EVOLVE: Self-Evolution of language model reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32752–32798, Suzhou, China. Association for Computational Linguistics.

Omkar Thawakar, Shravan Venkatraman, Ritesh Thawkar, Abdelrahman Shaker, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Khan. 2025. Evolmm: Self-evolving large multimodal models with continuous rewards. *Preprint*, arXiv:2511.16672.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Qinsi Wang, Bo Liu, Tianyi Zhou, Jing Shi, Yueqian Lin, Yiran Chen, Hai Helen Li, Kun Wan, and Wentian Zhao. 2025a. Vision-zero: Scalable vlm self-improvement via strategic gamified self-play. *Preprint*, arXiv:2509.25541.

Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Taha Kass-Hout, Furong Huang, and Cao Xiao. 2025b. Enhancing visual-language modality alignment in large vision language models via self-improvement. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 268–282, Albuquerque, New Mexico. Association for Computational Linguistics.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

Renqiu Xia, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Botian Shi, Junchi Yan, and Bo Zhang. 2025. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *IEEE Transactions on Image Processing*.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025. From generic empathy to personalized emotional support: A self-evolution framework for user preference alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18826–18853, Suzhou, China. Association for Computational Linguistics.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. Demystifying long chain-of-thought reasoning in llms. *Preprint*, arXiv:2502.03373.

Xiangning Yu, Zhuohan Wang, Linyi Yang, Haoxuan Li, Anjie Liu, Xiao Xue, Jun Wang, and Mengyue Yang. 2025. Causal sufficiency and necessity improves chain-of-thought reasoning. *Preprint*, arXiv:2506.09853.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Juntian Zhang, Song Jin, Chuanqi Cheng, Yuhan Liu, Yankai Lin, Xun Zhang, Yufei Zhang, Fei Jiang, Guojun Yin, Wei Lin, and Rui Yan. 2025a. Viper: Empowering the self-evolution of visual perception abilities in vision-language model. *Preprint*, arXiv:2510.24285.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2024a. Lmms-eval: Reality check on the evaluation of large multimodal models. *Preprint*, arXiv:2407.12772.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024b. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2025b. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1662, Vienna, Austria. Association for Computational Linguistics.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. 2024c. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Preprint*, arXiv:2406.09136.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024d. Multimodal chain-of-thought reasoning in language models. *Preprint*, arXiv:2302.00923.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024e. Multimodal chain-of-thought reasoning in language models. *Preprint*, arXiv:2302.00923.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. 2024. Calibrated self-rewarding vision language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Adam Zweiger, Jyothish Pari, Han Guo, Ekin Akyürek, Yoon Kim, and Pulkit Agrawal. 2025. Self-adapting language models. *Preprint*, arXiv:2506.10943.