MAGNET: Multi-Agent Collaboration with Confidence Filtering for Answer-Free Multimodal Reasoning

Anonymous ACL submission

Abstract

While supervised fine-tuning advances multimodal large language models (MLLMs), enhancing their reasoning remains constrained by the scarcity of high-quality Chain-of-Thought (CoT) data. Furthermore, distilling reasoning chains from powerful teacher models often suffers from a "learnability gap", as their complex patterns exceed smaller models' capacities, while purely self-supervised approaches are highly susceptible to reward hacking via visual shortcuts. To address this dilemma, we propose MAGNET (Multi-Agent Collaboration with Confidence Filtering), an answer-free preference alignment framework that leverages unlabeled data to boost intrinsic reasoning. MAGNET elicits diverse reasoning paths via confidence-guided decoding from the target model itself. To improve trajectory reliability, we introduce a progressive verification mechanism. We first perform outcome-level screening based on answer confidence. Subsequently, to mitigate visual shortcuts and multimodal hallucinations, we employ a text-only multi-agent calibration mechanism. By decoupling visual perception from evaluation, it provides fine-grained, process-level logical scrutiny while shielding verification from visual interference. Finally, the consensus trajectories and high-confidence rejected alternatives are formulated into contrastive preference pairs for Direct Preference Optimization (DPO). Extensive experiments on M^3CoT, ScienceQA, and GeoQA demonstrate that MAGNET improves accuracy in answer-free settings and enhances reasoning quality, as evidenced by substantial gains in Process Reward Model (PRM) scores.

1 Introduction

In recent years, Multimodal Large Language Models (MLLMs) (Liu et al., 2023; Achiam et al., 2023; Bai et al., 2023b) have advanced language understanding by integrating visual perception, achieving success in tasks like visual question answering and image captioning (Zhu et al., 2023; Zhang et al., 2024; Yin et al., 2024). However, as the paradigm shifts from perceptual understanding to structured reasoning, MLLMs must not only "see" visual inputs but also perform multi-step logical deduction and decision-making within complex visual-textual contexts (Wang et al., 2024b; Yang et al., 2023; Shi et al., 2024), posing a substantial challenge to their intrinsic reasoning capacities.

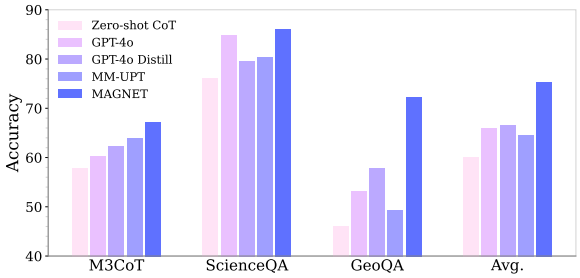

Figure 1: Accuracy comparison of different answer-free methods on three datasets. All methods use greedy decoding except GPT-4o, which adopts top-p sampling. MAGNET achieves a significant improvement over its Zero-shot baseline and outperforms all other methods.

To bolster the reasoning and decision-making capabilities of MLLMs, current post-training paradigms predominantly rely on three methodologies: supervised fine-tuning (SFT) on human-annotated reasoning data (Ouyang et al., 2022), reinforcement learning from human feedback (RLHF) (Bai et al., 2022), and reinforcement learning with verifiable outcomes (RLVR) (Guo et al., 2025). Despite their effectiveness, these frameworks face a bottleneck: the extreme scarcity and prohibitive annotation costs of high-quality, step-by-step Chain-of-Thought data. To exploit abundant unlabeled data, a straightforward alternative is off-policy distillation, which extracts reasoning chains from proprietary teacher models (e.g., GPT-4o) as pseudo-supervision. However, recent literature highlights a prominent *learnability gap* (Li et al., 2025): smaller student models frequently struggle to assimilate the extended and intricate reasoning traces distilled from larger counterparts,

as these trajectories contain implicit, multi-layered cross-modal assumptions that exceed the parameter scale and reasoning capacity of smaller models, as empirically demonstrated in Figure 1. Conversely, purely self-supervised techniques utilize online RL frameworks to elicit on-policy trajectories from the model's own distribution (Wei et al., 2025). Nevertheless, these methods heavily rely on coarse-grained, outcome-level rewards such as majority voting. In multimodal scenarios, this inevitably triggers severe *reward hacking*: models frequently exploit visual shortcuts or fabricate hallucinated logic to guess the correct final answer, thereby receiving unwarranted positive rewards despite fundamentally flawed reasoning. Moreover, integrating multi-agent verification into online RL loops causes prohibitive overhead and training instability.

These challenges motivate a critical question:

> ***How can we construct high-quality, process-level learnable self-training signals from unlabeled data to advance MLLM reasoning without ground-truth answers?***

To address this, we must resolve three challenges: **(i)** How can we elicit diverse and capacity-aligned reasoning trajectories without relying on ground-truth answers or teacher models that induce a learnability gap? **(ii)** How can we design a fine-grained verification mechanism that prevents reward hacking without ground-truth answers, especially when visual contexts introduce shortcuts and hallucinations? **(iii)** How can we efficiently transform process-level verification outcomes into optimization signals without the prohibitive computational latency and instability of online RL loops?

To address these challenges, we propose MAGNET (Multi-Agent Collaboration with Confidence Filtering), an answer-free preference alignment framework tailored for enhancing MLLM reasoning. MAGNET constructs a robust pipeline comprising three sequential stages: *Internal Activation*, *External Synergy*, and *Preference Alignment*. To avoid API-generated off-policy distillation, the **Internal Activation** stage leverages the pre-trained MLLM to generate multiple reasoning trajectories on unlabeled data. Confidence-guided decoding then selects high-quality candidate paths that activate the model's own reasoning capacity and better match its parameter scale and decoding distribution, thereby mitigating the learnability gap. To overcome the rampant reward hacking and single-model biases in traditional self-training, the **External Synergy** stage introduces a *progressive verification mechanism*. Following an initial outcome-level screening guided by answer confidence, we deploy an expert committee composed of multiple LLM agents to collaboratively evaluate the logical consistency of candidate trajectories from diverse perspectives. Crucially, by intentionally decoupling visual perception from logical evaluation, our text-only multi-agent committee is forced to provide meticulous, process-level structural scrutiny based entirely on textual coherence and deductive rigor. This visual-semantic decoupling repurposes stronger LLMs as text-only verifiers rather than off-policy generators, injecting fine-grained process-level signals while mitigating individual biases, visual shortcuts, and multimodal hallucinations. Finally, in the **Preference Alignment** stage, we formalize optimization as an offline alignment problem that reuses verified judgments as preference pairs, avoiding repeated expert calls and instability in online RL loops. By applying DPO to the consensus trajectories and high-confidence rejected alternatives, we optimize model preferences, introducing explicit negative penalty signals that precisely suppress hallucinated paths and shape superior reasoning behaviors in answer-free scenarios.

Our main contributions are three-fold:

❶ ***Answer-Free Preference Alignment Paradigm***. We propose MAGNET, an answer-free *generate-verify-align* framework that circumvents the distillation learnability gap and online RL reward hacking, unlocking MLLMs' intrinsic reasoning potential without ground-truth annotations.

❷ ***Visual-Semantic Decoupled Logical Verification***. We introduce a text-only multi-agent calibration mechanism. By isolating visual perception from logical evaluation, it filters out visual shortcuts and hallucinations, yielding high-fidelity, process-level pseudo-supervision.

❸ ***Offline Preference Alignment from Verified Signals***. We recast verification outcomes into contrastive DPO pairs, offering a stable, compute-efficient alternative to online RL. Extensive experiments show that MAGNET improves both zero-shot accuracy and PRM scores.

2 Related Work

Multimodal Large Language Models. Recent advancements in MLLMs mainly rely on two types of training data: converting vision-language datasets into instruction-tuning formats (Dai et al., 2023;

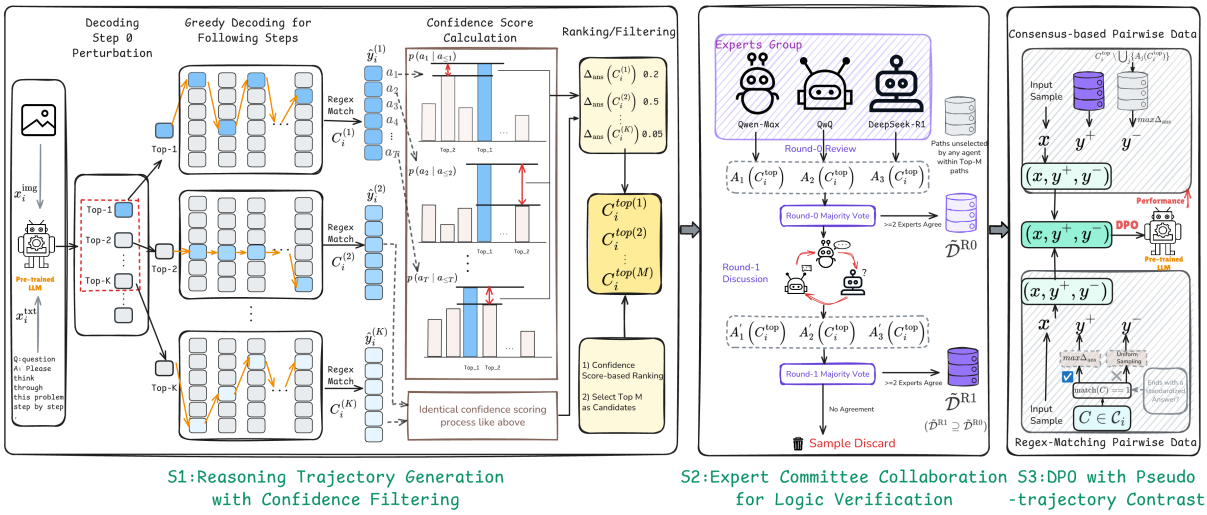

Figure 2: Overview of the MAGNET framework. MAGNET is a three-stage answer-free preference alignment framework, consisting of confidence-based trajectory generation, text-only multi-agent collaborative verification, and Direct Preference Optimization (DPO). By intentionally decoupling visual perception during evaluation, it improves the logical consistency and reasoning capability of the model in answer-free settings.

Liu et al., 2023, 2024a), and generating synthetic supervision with powerful proprietary models via multi-turn prompting (Achiam et al., 2023; Liu et al., 2024b; Wu et al., 2024). Genixer (Zhao et al., 2024) takes a step further by enabling MLLMs to self-generate visual instruction data. While these approaches have significantly advanced multimodal understanding, they predominantly depend on external annotations or ground-truth guidance. In contrast, our work targets the reasoning aspect explicitly, exploring how to leverage internally generated trajectories and external agent-based collaboration to enhance path-level decision quality without relying on ground-truth answers.

Answer-Free Preference Alignment. Recent studies have begun to explore methods for improving MLLMs under fully unlabeled settings. Common strategies include consistency-based mechanisms such as self-play, multi-path voting, and confidence-based selection (Yu et al., 2025; Zhou et al., 2024; Ahn et al., 2025). For example, Absolute Zero (Zhao et al., 2025) introduces a "self-ask–self-answer–self-evaluate" paradigm to enable zero-shot self-improvement in language models. In the multimodal domain, ReVisual-R1 (Chen et al., 2025) integrates curriculum learning with preference modeling, while SENA (Tan et al., 2025) and STIC (Deng et al., 2024) utilize visual cues to construct structured preference pairs that guide model optimization.

MM-UPT (Wei et al., 2025) applies GRPO (Shao et al., 2024) for unsupervised optimization of MLLMs, constructing pseudo-reward signals via majority voting over multiple sampled responses. While MM-UPT demonstrates the effectiveness of outcome-based consensus, it primarily focuses on final answer agreement. Building upon these foundations, MAGNET introduces a complementary perspective by focusing on fine-grained, process-level logical verification. By deliberately employing a text-only multi-agent committee, we decouple visual perception from logical evaluation, providing rigorous, step-level feedback to complement outcome-level consensus, thereby achieving high-fidelity reasoning enhancement without ground-truth annotations.

3 Method

Problem Definition. Given an unlabeled multimodal dataset $\mathcal{D} = \{x_i\}_{i=1}^{N}$, each sample $x_i = (x_i^{\text{img}}, x_i^{\text{txt}}, \mathcal{O}_i)$ lacks a ground-truth answer. A pre-trained MLLM $\mathcal{M}_0$ generates diverse reasoning trajectories $\mathcal{C}_i = \{C_i^{(1)}, \ldots, C_i^{(K)}\}$, where each path $C_i^{(k)}$ concludes with a predicted option $\hat{y}_i^{(k)} \in \mathcal{O}_i$. Our goal is to optimize the model's preference over $\mathcal{C}_i$ in a strictly answer-free setting. The core challenge is extracting reliable preference pairs (y_i^+, y_i^-) from this noisy pool by mitigating visual shortcuts and evaluating logical rigor without human intervention.

3.1 Framework Overview

Answer-free multimodal reasoning faces a critical dilemma: the learnability gap induced by teacher distillation, and reward hacking prevalent in purely self-supervised methods. To address this, MAGNET introduces a three-stage preference alignment

framework, as illustrated in Figure 2. It begins by leveraging confidence-guided CoT decoding from the target model itself to construct capacity-aligned reasoning paths. Then, an expert committee of text-only LLM agents collaboratively evaluates these paths. By explicitly decoupling visual perception from the evaluation process, it enforces pure logical scrutiny and effectively mitigates visual shortcuts. Finally, the model is optimized via Direct Preference Optimization (DPO) (Rafailov et al., 2023) over contrastive preference pairs. This framework forms a highly robust loop of generation, decoupled verification, and alignment, unlocking MLLMs' intrinsic reasoning potential without human-annotated data.

3.2 Reasoning Trajectory Generation with Confidence Filtering

Our framework operates under the fundamental premise that contemporary pre-trained MLLMs possess latent reasoning capacities. However, these capabilities are often suppressed by conventional greedy decoding (Wan et al., 2025; Romera-Paredes et al., 2024). To unlock this potential, we employ CoT decoding. Unlike traditional prompting techniques that impose handcrafted human inductive biases, CoT decoding provides a lightweight, prompt-free mechanism to naturally elicit diverse reasoning paths from the model's intrinsic manifold. **Rather than relying on off-policy distillation from proprietary teachers, we sample paths from the model's own distribution.** This generation is performed offline prior to DPO to ensure computational efficiency, and by sampling from the model's own distribution, it guarantees strict capacity alignment and effectively circumvents the learnability gap (Li et al., 2025).

Given a multimodal input $(x_i^{\text{img}}, x_i^{\text{txt}})$, we implement this decoding strategy by applying top-k sampling on the first decoded token to promote diversity, followed by greedy decoding. This yields candidate paths $\mathcal{C}_i = \{C_i^{(k)}\}_{k=1}^{K}$. Each path $C_i^{(k)}$ concludes with a predicted answer $\hat{y}_i^{(k)}$, extracted via regex. We apply a binary indicator $\texttt{match}(C_i^{(k)}) \in \{0, 1\}$ to filter out structurally incomplete paths. To assess the reliability of the valid trajectories, we compute an answer confidence score Δ_{ans} based on the average probability margin between the top-1 and top-2 tokens across the answer span $\hat{y}_i^{(k)}$:

$$\Delta_{\text{ans}}(C_i^{(k)}) = \frac{1}{|\hat{y}_i^{(k)}|} \sum_{a_t \in \hat{y}_i^{(k)}} \left(p(a_t^{(1)} \mid a_{<t}) - p(a_t^{(2)} \mid a_{<t}) \right) \quad (1)$$

Algorithm 1 MAGNET

Input: x_i, $\mathcal{M}_0$, agents $\mathcal{A}_{1:E}$; **Output:** $\mathcal{M}$
$\mathcal{C}_i \leftarrow \text{CoT}(\mathcal{M}_0, x_i)$; $\mathcal{T}_i \leftarrow \text{Top}_M(\mathcal{C}_i, \Delta_{\text{ans}})$
$\mathcal{V}_0 \leftarrow \{\mathcal{A}_j(\mathcal{T}_i)\}_{j=1}^{E}$
if $\text{Maj}(\mathcal{V}_0)$ exists **then**
  $y_i^+ \leftarrow \text{Maj}(\mathcal{V}_0)$; $\mathcal{V}_i \leftarrow \mathcal{V}_0$
else
  $\mathcal{V}_1 \leftarrow \text{Delib}(\mathcal{V}_0)$; discard if no majority
  $y_i^+ \leftarrow \text{Maj}(\mathcal{V}_1)$; $\mathcal{V}_i \leftarrow \mathcal{V}_0 \cup \mathcal{V}_1$
end if
$y_i^- \leftarrow \text{HardNeg}(\mathcal{T}_i, \mathcal{V}_i, \Delta_{\text{ans}})$
$(y_{i,\text{reg}}^+, y_{i,\text{reg}}^-) \leftarrow \text{RegexPair}(\mathcal{C}_i)$
$\mathcal{P}_i \leftarrow \{(y_i^+, y_i^-), (y_{i,\text{reg}}^+, y_{i,\text{reg}}^-)\}$
$\mathcal{M} \leftarrow \text{DPO}(\mathcal{M}_0, x_i, \mathcal{P}_i)$

where $a_t^{(1)}$ and $a_t^{(2)}$ denote the top-1 and top-2 predictions at position t. We rank all valid paths by Δ_{ans} and retain the top M to form the candidate set $\mathcal{C}_i^{\text{top}}$. This confidence-based screening serves as the first line of defense by filtering out paths with severe visual perception failures, ensuring that the retained candidates are grounded in visual facts before entering text-only logical verification.

3.3 Expert Committee Collaboration for Logic Verification

While the aforementioned confidence filtering guarantees that the candidate set $\mathcal{C}_i^{\text{top}}$ is largely grounded in visual facts, it does not ensure the absence of logical flaws or reasoning leaps. To rigorously assess process-level structural coherence, we introduce a text-only, multi-agent collaborative verification framework. We engage $E = 3$ API-based LLM experts $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ (Qwen-Max (Bai et al., 2023a), QwQ (Yang et al., 2025), and DeepSeek-R1 (Guo et al., 2025)) to independently select the most logically consistent trajectory from $\mathcal{C}_i^{\text{top}}$.

Why Text-Only Experts instead of MLLM Judges? A natural question arises: why not employ powerful proprietary MLLMs (e.g., GPT-4o) as verifiers? We intentionally blind the expert committee to the image due to two critical considerations. First, *Confirmation Bias in Visual Shortcuts*. If a verifier has access to the image, it is highly susceptible to the same multimodal hallucinations and spurious correlations that the generator experiences. An MLLM judge might easily endorse a flawed reasoning trajectory simply because it observes the same misleading visual cues, failing to critically examine the deduction steps. By decoupling visual perception from evaluation, we force the committee to rely strictly on logical scrutiny. Second, *Data Contamination and Fair Evaluation*. Contemporary closed-source MLLMs are extensively trained on standard multimodal benchmarks.

Utilizing them as image-aware judges risks introducing data contamination, where the verifier relies on its memorized knowledge of the specific dataset rather than evaluating the intrinsic quality of the generated trajectory.

Multi-Agent Debate. Each expert $\mathcal{A}_j$ receives the question and text-only candidates to select a preferred path $C_{i,j}^{(0)} = \mathcal{A}_j(\mathcal{C}_i^{\text{top}}) \in \mathcal{C}_i^{\text{top}}$ with a textual justification. These form the Round-0 vote set $\mathcal{V}_0 = \{C_{i,j}^{(0)}\}_{j=1}^{E}$. If a majority agreement ($\geq$ 2 experts) is reached, we set the consensus trajectory as $C_i^+ = \text{MajorityVote}(\mathcal{V}_0)$, define $\mathcal{V}_i = \mathcal{V}_0$, and add it to the Round-0 dataset, $\tilde{\mathcal{D}}^{\text{R0}} \leftarrow \tilde{\mathcal{D}}^{\text{R0}} \cup \{(x_i, C_i^+)\}$. If all experts disagree, a Round-1 debate is initiated. Experts review peer selections and rationales from Round-0, prompting them to critique the alternatives and optionally revise their own votes. Let $\mathcal{V}_1 = \{C_{i,j}^{(1)}\}_{j=1}^{E}$ denote the revised Round-1 vote set. If a majority emerges after discussion, we set $C_i^+ = \text{MajorityVote}(\mathcal{V}_1)$, define $\mathcal{V}_i = \mathcal{V}_0 \cup \mathcal{V}_1$, and append the agreed path to the final dataset $\tilde{\mathcal{D}}^{\text{R1}} \leftarrow \tilde{\mathcal{D}}^{\text{R1}} \cup \{(x_i, C_i^+)\}$, where $\tilde{\mathcal{D}}^{\text{R1}} \supseteq \tilde{\mathcal{D}}^{\text{R0}}$; otherwise, the sample is discarded. In practice, fewer than 0.1% of samples fail to reach consensus after the debate. This mechanism mitigates individual biases, blocks visual hallucinations, and yields high-fidelity pseudo-labels without ground-truth annotations.

3.4 Direct Preference Optimization over Contrastive Trajectories

To efficiently transform process-level verification outcomes into stable optimization signals, we incorporate Direct Preference Optimization (DPO) (Rafailov et al., 2023) over the verified trajectories. A natural alternative would be online reinforcement learning methods such as GRPO (Shao et al., 2024), which have recently been applied to multimodal reasoning (Wei et al., 2025). However, online RL presents two practical obstacles in our setting. First, it requires repeated reward queries during training; in our framework, each reward query would entail a full round of expert committee calls, making the cost prohibitive. Second, online exploration introduces training instability, particularly when the reward signal itself is a learned pseudo-label rather than a ground-truth verifier. In contrast, our offline DPO formulation reuses a single round of committee judgments, decoupling verification cost from training iterations. By explicitly contrasting highly reliable preference pairs (x_i, y_i^+, y_i^-), DPO provides stable, compute-efficient preference alignment. These pairs are constructed from two complementary answer-free signals: agent consensus and structural compliance.

Consensus-based Pairwise Data. We leverage the outcome of the expert committee to define strong logical preference supervision. Specifically, for each retained sample in $\tilde{\mathcal{D}}^{\text{R1}}$, we denote its consensus trajectory as C_i^+ and treat it as the positive sample y_i^+. For the hard-negative counterpart $y_{i,\text{cons}}^-$, we select the highest-confidence path among the remaining candidates that were *not* selected by any expert during either Round-0 or Round-1:

$$\begin{aligned} y_{i,\text{cons}}^+ &= C_i^+, \\ y_{i,\text{cons}}^- &= \arg\max_{C \in \mathcal{C}_i^{\text{top}} \setminus \mathcal{V}_i} \Delta_{\text{ans}}(C). \end{aligned} \tag{2}$$

This design guarantees that the negative path is logically distinct from expert-approved trajectories, yet still competitive in answer confidence—thus forming a hard-negative pair. Such construction enhances the model's ability to distinguish subtle logical flaws and hallucinations, aligning its generation behavior with expert-endorsed rigorous logic.

Regex-Matching Pairwise Data. To simultaneously strengthen format compliance and prompt adherence, we introduce a secondary set of preference pairs based on structural output validity. Let $\mathcal{M}_i = \{C \in \mathcal{C}_i \mid \texttt{match}(C) = 1\}$ denote the set of paths ending with a well-formed answer conclusion. We define the structurally preferred and dispreferred samples as:

$$\begin{aligned} y_{i,\text{reg}}^+ &= \arg\max_{C \in \mathcal{M}_i} \Delta_{\text{ans}}(C), \\ y_{i,\text{reg}}^- &\sim \mathcal{U}\left(\{C \in \mathcal{C}_i \mid \texttt{match}(C) = 0\}\right), \end{aligned} \tag{3}$$

where the unmatched paths $y_{i,\text{reg}}^-$ are decodable but structurally incomplete or ill-formed. This contrast encourages the model to prefer outputs that not only reason correctly but also yield interpretable and strictly formatted conclusions.

Optimization Objective. Both consensus-based and regex-based preference pairs are collected into $\mathcal{D}_{\text{pref}}$ and optimized jointly using the standard DPO loss (Rafailov et al., 2023) (full formulation in Appendix). By integrating this dual-source optimization, MAGNET systematically reshapes the model's preference toward reasoning trajectories that are simultaneously *logically faithful* and *structurally complete*, achieving robust reasoning alignment in answer-free scenarios.

Strategy	CoT	Label	M^{3}CoT		ScienceQA		GeoQA		Avg.	
			Greedy	Agg.	Greedy	Agg.	Greedy	Agg.	Greedy	Agg.
SFT w/ label										
QA	×	✓	41.03$^{\downarrow 16.8}$	40.29$^{\downarrow 27.0}$	52.78$^{\downarrow 23.4}$	50.98$^{\downarrow 33.7}$	32.10$^{\downarrow 13.9}$	33.55$^{\downarrow 37.0}$	41.97$^{\downarrow 18.0}$	41.61$^{\downarrow 32.6}$
CoT	✓	✓	68.51$^{\uparrow 10.7}$	73.17$^{\uparrow 5.9}$	90.96$^{\uparrow 14.8}$	93.14$^{\uparrow 8.5}$	81.03$^{\uparrow 35.0}$	85.41$^{\uparrow 14.9}$	80.17$^{\uparrow 20.2}$	83.91$^{\uparrow 9.7}$
Training-Free										
Zero-shot QA	×	×	58.46$^{\uparrow 0.6}$	–	69.34$^{\downarrow 6.8}$	–	53.85$^{\uparrow 7.8}$	–	60.55$^{\uparrow 0.6}$	–
Zero-shot CoT	✓	×	57.81	67.28	76.14	84.65	46.02	70.56	59.99	74.16
GPT-4o[a]	×	×	60.15$^{\uparrow 2.3}$	–	84.77$^{\uparrow 8.6}$	–	53.02$^{\uparrow 7.0}$	–	65.98$^{\uparrow 6.0}$	–
SFT w/ pseudo-label										
GPT-4o Distill	×	×	62.35$^{\uparrow 4.5}$	63.53$^{\downarrow 3.8}$	79.45$^{\uparrow 3.3}$	86.00$^{\uparrow 1.3}$	57.81$^{\uparrow 11.8}$	63.26$^{\downarrow 7.3}$	66.54$^{\uparrow 6.5}$	70.93$^{\downarrow 3.2}$
Aggregation Select	✓	×	57.42$^{\downarrow 0.4}$	62.08$^{\downarrow 5.2}$	77.02$^{\uparrow 0.9}$	80.17$^{\downarrow 4.5}$	65.70$^{\uparrow 19.7}$	70.95$^{\uparrow 0.4}$	66.71$^{\uparrow 6.7}$	71.07$^{\downarrow 3.1}$
DeepSeek-R1 Select	✓	×	59.53$^{\uparrow 1.7}$	64.54$^{\downarrow 2.7}$	84.93$^{\uparrow 8.8}$	87.32$^{\uparrow 2.7}$	67.24$^{\uparrow 21.2}$	73.87$^{\uparrow 3.3}$	70.57$^{\uparrow 10.6}$	75.24$^{\uparrow 1.1}$
Qwen-Max Select	✓	×	61.48$^{\uparrow 3.7}$	63.81$^{\downarrow 3.5}$	84.75$^{\uparrow 8.6}$	85.73$^{\uparrow 1.1}$	70.82$^{\uparrow 24.8}$	76.79$^{\uparrow 6.2}$	72.35$^{\uparrow 12.4}$	75.44$^{\uparrow 1.3}$
QwQ Select	✓	×	51.55$^{\downarrow 6.3}$	64.84$^{\downarrow 2.4}$	85.62$^{\uparrow 9.5}$	87.75$^{\uparrow 3.1}$	71.35$^{\uparrow 25.3}$	76.92$^{\uparrow 6.4}$	69.51$^{\uparrow 9.5}$	76.50$^{\uparrow 2.3}$
Answer-Free Post-Training										
MM-UPT	✓	×	63.81$^{\uparrow 6.0}$	69.54$^{\uparrow 2.3}$	80.23$^{\uparrow 4.1}$	86.96$^{\uparrow 2.3}$	49.20$^{\uparrow 3.2}$	67.37$^{\downarrow 3.2}$	64.41$^{\uparrow 4.4}$	74.62$^{\uparrow 0.5}$
Ours	✓	×	**67.17**$^{\uparrow 9.4}$	**70.10**$^{\uparrow 2.8}$	**86.08**$^{\uparrow 9.9}$	**87.93**$^{\uparrow 3.3}$	**72.28**$^{\uparrow 26.3}$	**83.82**$^{\uparrow 13.3}$	**75.18**$^{\uparrow 15.2}$	**80.62**$^{\uparrow 6.5}$

[a]Uses top-p sampling instead of greedy decoding.

Table 1: Accuracy comparison across strategies. Underlined: Zero-shot CoT baseline. **Bold**: Best performance among answer-free methods. ↑/↓: Relative change from the baseline.

4 Experiments

4.1 Experimental Setup

Datasets and Baselines. We evaluate on three multimodal reasoning benchmarks: M^{3}CoT (Chen et al., 2024), ScienceQA (Lu et al., 2022), and GeoQA (Chen et al., 2021) (details in Appendix). Baselines include zero-shot inference and SFT with human annotations (w/ and w/o CoT). For answer-free baselines, we compare against *GPT-4o distillation*, single-agent path selection (*DeepSeek-R1/Qwen-Max/QwQ Select*), and *Aggregation Select* (Wang and Zhou, 2024). We include *MM-UPT* (Wei et al., 2025), a state-of-the-art GRPO-based framework for answer-free multimodal reasoning. This diverse suite ensures rigorous evaluation of MAGNET under answer-free settings.

Evaluation Metrics (Accuracy). We report two accuracy metrics (omitting % for clarity): **Greedy** denotes the top-1 decoding path, $\hat{y}_i^{\text{greedy}} = \hat{y}_i^{(1)}$, **Agg.** denotes the answer selected by aggregating multiple reasoning paths using answer confidence, $\hat{y}_i^{\text{agg}} = \arg\max_{y \in \mathcal{O}_i} \sum_{k:\, \hat{y}_i^{(k)} = y} \Delta_{\text{ans}}(C_i^{(k)})$. As MAGNET explicitly elicits diverse reasoning paths, it naturally aligns with the Agg. metric. This effectively captures the model's test-time scaling potential, unlocking stronger test-time reasoning performance than single greedy decoding.

Evaluation Metrics (PRM). While outcome-based accuracy strictly measures the correctness of the final answer, it fails to capture whether the model relied on spurious visual shortcuts or reward hacking to guess that answer. To rigorously quantify the *process-level* reasoning quality, we introduce the Process Reward Model (PRM) score as a novel, complementary metric. Specifically, we employ VisualPRM-8B (Wang et al., 2025) to assign a scalar score to each reasoning trajectory. This score evaluates whether the rationale exhibits a coherent step-by-step logical structure, derives the final answer through valid intermediate deductions, and avoids reasoning jumps. It is crucial to note that the PRM score is completely isolated from our optimization pipeline; it serves purely as an external diagnostic tool to monitor intrinsic reasoning improvements across different stages.

Implementation Details. Implemented via LLaMA-Factory (Zheng et al., 2024) on 8×A100 (80G) GPUs, we adopt Qwen2.5-VL-7B-Instruct (Wang et al., 2024a) (8k context) as the backbone. We perform full-parameter fine-tuning during both SFT and DPO to ensure deep cross-modal representational alignment. To prevent catastrophic forgetting, we employ a conservative optimization setup: SFT and DPO span 2 and 1 epochs respectively, with a batch size of 8 and a cosine learning rate (0.1 warm-up) peaking at an ultra-low 5×10^{-7}. For DPO, the KL penalty is $\beta = 0.2$, and an auxiliary SFT loss ($\mu = 1.0$) is added to strictly maintain format compliance. Memory efficiency is guaranteed via `bf16` precision, FlashAttention-V2 (Dao, 2023), and DeepSpeed ZeRO-3 (Rasley et al., 2020). During trajectory generation, we sample $K = 10$ paths and retain the top $M = 5$ by answer confidence and conclusion uniqueness. **This keeps API-based verification controllable: experts evaluate text-only top-M candidates offline once, balancing exploration with verification overhead.**

4.2 Empirical Results

Accuracy gains without answers. We examine whether MAGNET can improve reasoning with-

Strategy	CoT	Label	M^{3}CoT		ScienceQA		GeoQA		Avg.	
			Greedy	Agg.	Greedy	Agg.	Greedy	Agg.	Greedy	Agg.
SFT w/ label										
CoT	✓	✓	88.66$^{\uparrow 26.1}$	88.65$^{\uparrow 25.5}$	93.08$^{\uparrow 20.8}$	93.56$^{\uparrow 20.6}$	80.65$^{\uparrow 23.7}$	80.24$^{\uparrow 22.9}$	87.46$^{\uparrow 23.5}$	87.48$^{\uparrow 23.0}$
Training-Free										
Zero-shot CoT	✓	×	62.55	63.11	72.32	72.97	56.95	57.38	63.94	64.49
SFT w/ pseudo-label										
Aggregation Select	✓	×	79.61$^{\uparrow 17.1}$	85.07$^{\uparrow 22.0}$	88.54$^{\uparrow 16.2}$	91.19$^{\uparrow 18.2}$	60.47$^{\uparrow 3.5}$	79.15$^{\uparrow 21.8}$	76.21$^{\uparrow 12.3}$	85.14$^{\uparrow 20.7}$
DeepSeek-R1 Select	✓	×	52.87$^{\downarrow 9.7}$	53.96$^{\downarrow 9.2}$	66.78$^{\downarrow 5.5}$	67.32$^{\downarrow 5.6}$	58.21$^{\uparrow 1.3}$	58.96$^{\uparrow 1.6}$	59.29$^{\downarrow 4.7}$	60.08$^{\downarrow 4.4}$
Qwen-Max Select	✓	×	62.33$^{\downarrow 0.2}$	63.35$^{\uparrow 0.2}$	70.73$^{\downarrow 1.6}$	70.74$^{\downarrow 2.2}$	82.34$^{\uparrow 25.4}$	81.81$^{\uparrow 24.4}$	71.80$^{\uparrow 7.9}$	71.97$^{\uparrow 7.5}$
QwQ Select	✓	×	53.94$^{\downarrow 8.6}$	54.94$^{\downarrow 8.2}$	68.37$^{\downarrow 4.0}$	69.51$^{\downarrow 3.5}$	63.05$^{\uparrow 6.1}$	64.82$^{\uparrow 7.4}$	61.79$^{\downarrow 2.2}$	63.09$^{\downarrow 1.4}$
Answer-Free Post-Training										
MM-UPT	✓	×	89.75$^{\uparrow 27.2}$	91.24$^{\uparrow 28.1}$	92.46$^{\uparrow 20.1}$	94.87$^{\uparrow 21.9}$	73.07$^{\uparrow 16.1}$	73.38$^{\uparrow 16.0}$	85.09$^{\uparrow 21.2}$	86.50$^{\uparrow 22.0}$
Ours	✓	×	**91.53**$^{\uparrow 29.0}$	**91.42**$^{\uparrow 28.3}$	**94.60**$^{\uparrow 22.3}$	**94.97**$^{\uparrow 22.0}$	**81.68**$^{\uparrow 24.7}$	**82.80**$^{\uparrow 25.4}$	**89.27**$^{\uparrow 25.3}$	**89.73**$^{\uparrow 25.2}$

Table 2: Comparison of PRM scores across different strategies. Formatting conventions follow Table 1.

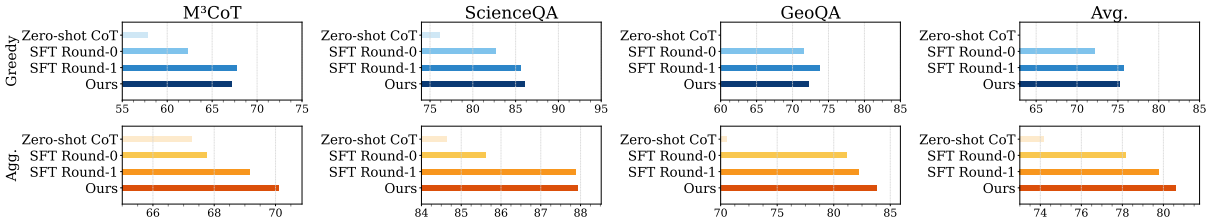

Figure 3: Accuracy improvements across different stages of our multi-round no-answer training framework. Each stage reflects a key step in pseudo-label refinement and preference optimization.

out ground-truth answers or human-written rationales. Table 1 reports two complementary metrics: Greedy accuracy measures the top-1 decoding path, while Agg. accuracy evaluates whether the model benefits from multiple internally generated reasoning trajectories. Compared with Zero-shot CoT, MAGNET improves average accuracy from 59.99 to 75.18 under Greedy decoding and from 74.16 to 80.62 under Agg., yielding gains of +15.19 and +6.46 points, respectively. These improvements are consistent on M^{3}CoT, ScienceQA, and GeoQA, showing that the gains are not tied to a single benchmark. Although supervised CoT SFT remains a strong reference by using gold answers and human-written rationales, MAGNET substantially narrows the gap: its average Agg. accuracy reaches 80.62, only 3.29 points below supervised CoT SFT. This suggests that high-quality answer-free preference signals can recover a large portion of the benefit usually brought by annotated reasoning data.

Comparison with answer-free methods. Among answer-free methods, MAGNET achieves the strongest performance. It outperforms MM-UPT by +10.77 points under Greedy decoding and +6.00 points under Agg. MM-UPT shows that MLLMs can benefit from using internal majority voting as a self-judge, but such outcome-level self-reward is vulnerable to reward hacking in multimodal scenarios. In contrast, MAGNET introduces a decoupled text-only expert committee to avoid self-evaluation traps and impose stricter logical scrutiny. This leads to more reliable answer-free reward signals by combining confidence-based candidate filtering with fine-grained process-level verification.

Comparison with teacher distillation. Table 1 compares MAGNET with GPT-4o prompting and GPT-4o distillation. GPT-4o prompting obtains an average accuracy of 65.98, while GPT-4o distillation reaches 66.54 under Greedy and 70.93 under Agg. In contrast, MAGNET achieves 75.18 and 80.62, improving over distillation by +8.64 and +9.69 points, respectively. These results show that off-policy teacher supervision does not provide learnable signals for a smaller MLLM. Rather than imitating teacher-generated pseudo-labels, MAGNET constructs preference signals from trajectories generated by the target model itself, making the training data better aligned with the model's own reasoning capacity and decoding distribution.

Effectiveness of the multi-agent committee. MAGNET also outperforms single-agent selection baselines, including DeepSeek-R1 Select, Qwen-Max Select, and QwQ Select, as well as aggregation-based pseudo-labeling methods. These baselines rely on either answer-level aggregation or a single strong text LLM to select pseudo trajectories, making the final signal sensitive to a specific selection rule or verifier preference. In contrast, MAGNET aggregates complementary judgments from multiple text-only experts and further resolves disagreements through deliberation. The gains over Aggregation Select and single-agent selectors indicate that confidence filtering and committee verification are complementary: the former filters can-

didate paths by answer confidence, while the latter further checks logical coherence and suppresses brittle or hallucinated reasoning chains.

Process-level reasoning quality. Accuracy alone cannot determine whether a model reaches the correct answer through valid reasoning or shortcuts. We therefore report PRM scores as an external diagnostic of process-level reasoning quality. As shown in Table 2, MAGNET achieves average PRM scores of 89.27 under Greedy and 89.73 under Agg., improving over Zero-shot CoT by +25.33 and +25.24 points, respectively. It also exceeds MM-UPT by more than three points on average. Since the PRM is never used during training, these results indicate that the learned preferences improve not only final-answer accuracy but also the structural coherence of generated reasoning chains.

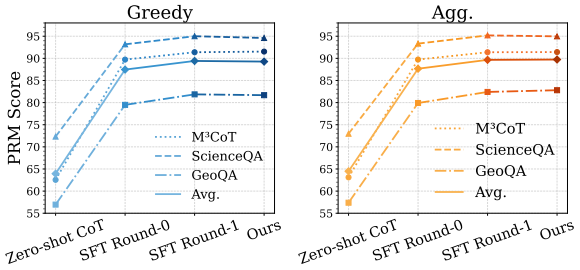

Figure 4: PRM score comparison across stages.

Effectiveness of visual-semantic decoupling. Text-only verification focuses on logic rather than visual perception. The large PRM gains support our decoupled design: after confidence filtering screens candidate trajectories based on answer confidence, text-only experts can safely focus on logical rigor without being directly exposed to visual shortcuts. MAGNET obtains higher PRM scores than supervised CoT SFT. This does not mean that answer-free training generally replaces human supervision; rather, it shows that explicitly constructing contrastive preference pairs from verified trajectories can encourage reasoning structures highly favored by an independent process-level evaluator.

Progressive signal refinement. Figures 3 and 4 show how answer-free signals are progressively refined across training stages. Starting from Zero-shot CoT, SFT on Round-0 consensus trajectories improves both accuracy and PRM, indicating that initial committee agreement provides useful pseudo trajectories. Round-1 deliberation further resolves ambiguous cases and stabilizes reasoning behavior. The post-hoc diagnosis in Table 3, where ground-truth answers are used only for analysis, further explains this trend. Answer-confidence filtering improves over raw Greedy selection, and committee

Strategy	M^{3}CoT	ScienceQA	GeoQA
Confidence-based filtering			
Greedy	59.06	74.57	46.61
CoT Confidence	48.85	67.65	20.58
Answer Confidence	67.96	85.61	70.42
Aggregation	68.94	86.25	72.48
Top-M path selection			
Top-1 Path	68.33	86.54	72.09
Top-2 Path	68.13	86.38	74.28
Top-3 Path	68.20	86.07	72.48
Top-4 Path	68.24	86.60	71.92
Top-5 Path	68.52	85.91	71.96
Single-agent verification			
Qwen-Max	70.05	87.90	80.31
QwQ	71.21	88.42	84.94
DeepSeek-R1	**71.45**	**88.72**	**85.08**
Committee consensus			
Round-0	70.97	88.57	84.46
Round-1	**71.21**	**88.61**	**84.94**

Table 3: Post-hoc Diagnostic Accuracy of Answer-Free Signals on the Training Candidate Pool.

consensus remains competitive with strong individual verifiers while reducing single-evaluator dependence. These results show that MAGNET gradually transforms noisy self-generated paths into more reliable, logically verified answer-free signals.

DPO effectiveness. The final DPO stage converts verified trajectories and high-confidence rejected candidates into contrastive preference pairs. Although Greedy accuracy may slightly fluctuate after DPO, Agg. accuracy continues to improve because DPO optimizes relative preferences over multiple trajectories rather than only the single top-1 output. Meanwhile, high PRM scores indicate that preference alignment preserves the reasoning structures established by expert verification. Together with the downstream results in Tables 1 and 2, this demonstrates that DPO effectively turns process-level verification outcomes into stable optimization signals for answer-free reasoning alignment.

5 Conclusion

This paper proposes MAGNET, an answer-free framework that improves MLLM reasoning through three complementary stages: self-generated trajectory sampling, text-only multi-agent logical verification, and offline DPO alignment. The key insight is that decoupling visual perception from logical evaluation—employing strong LLMs as text-only verifiers—yields process-level pseudo-supervision free from visual shortcuts. Experiments on M^{3}CoT, ScienceQA, and GeoQA demonstrate consistent improvements in accuracy and PRM scores, substantially narrowing the gap with fully supervised methods. Our results suggest that text-only verification is a practical paradigm for answer-free multimodal reasoning alignment.

6 Limitations

Visual grounding gap in text-only verification. By design, the expert committee evaluates reasoning trajectories using only textual content, intentionally decoupling visual perception from logical scrutiny. While this mitigates visual shortcuts and multimodal hallucinations, it also introduces a trade-off: a trajectory may appear logically coherent in text yet rely on incorrect visual grounding. Text-only verification is inherently insensitive to errors that require cross-modal alignment with the image. We note that the confidence-based filtering stage partially compensates by screening for answer-level visual consistency before text-only review.

API dependency and cost. MAGNET relies on API-based text LLMs as expert verifiers. Each sample incurs three Round-0 calls, with at most three additional calls for Round-1 deliberation (triggered for $< 0.1\%$ of samples). Verification is offline and performed once—not per training iteration—and text-only LLM APIs are substantially cheaper than multimodal alternatives. A cost breakdown is provided in the Appendix.

Evaluation scope. Our experiments are conducted on three multiple-choice multimodal reasoning benchmarks using Qwen2.5-VL-7B-Instruct as the backbone. Generalization to open-ended generation, broader benchmark categories, and diverse model scales remains to be validated. We use VisualPRM-8B as an external process-level diagnostic; since its scores primarily reflect textual reasoning coherence, the PRM improvements should be interpreted as evidence of enhanced structural reasoning quality rather than comprehensive multimodal reasoning fidelity.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Daechul Ahn, Yura Choi, San Kim, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. 2025. Isr-dpo: Aligning large multimodal models for videos by iterative self-retrospective dpo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1728–1736.

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024. M^3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*.

Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. 2025. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *arXiv preprint arXiv:2506.04207*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Y Zou, Kai-Wei Chang, and Wei Wang. 2024. Enhancing large vision language models with self-training on image comprehension. *Advances in Neural Information Processing Systems*, 37:131369–131397.

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*.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. 2025. Small models struggle to learn from strong reasoners. *arXiv preprint arXiv:2502.12143*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge, january 2024. *URL https://llava-vl. github. io/blog/2024-01-30-llava-next*, 1(8).

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

Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in neural information processing systems*, 35:2507–2521.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, and 1 others. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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*.

Wentao Tan, Qiong Cao, Yibing Zhan, Chao Xue, and Changxing Ding. 2025. Beyond human data: Aligning multimodal large language models by iterative self-evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7202–7210.

Fanqi Wan, Weizhou Shen, Shengyi Liao, Yingcheng Shi, Chenliang Li, Ziyi Yang, Ji Zhang, Fei Huang, Jingren Zhou, and Ming Yan. 2025. Qwenlong-l1: Towards long-context large reasoning models with reinforcement learning. *arXiv preprint arXiv:2505.17667*.

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

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, and 1 others. 2025. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. *Advances in Neural Information Processing Systems*, 37:66383–66409.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024b. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.

Lai Wei, Yuting Li, Chen Wang, Yue Wang, Linghe Kong, Weiran Huang, and Lichao Sun. 2025. Unsupervised post-training for multi-modal llm reasoning via grpo. *arXiv preprint arXiv:2505.22453*.

Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. 2024. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22227–22238.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, and 1 others. 2025. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19985–19995.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. 2025. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*.

Henry Hengyuan Zhao, Pan Zhou, and Mike Zheng Shou. 2024. Genixer: Empowering multimodal large language model as a powerful data generator. In *European Conference on Computer Vision*, pages 129–147. Springer.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pages 400–410.

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. *Advances in Neural Information Processing Systems*, 37:51503–51531.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Appendix

A Aggregation Select

We include AGGREGATION SELECT as an answer-free pseudo-labeling baseline following (Wang and Zhou, 2024). This method selects a pseudo-label by aggregating answer-level confidence over multiple generated reasoning trajectories. Different from MAGNET, which verifies the logical consistency of the full reasoning chain through text-only expert collaboration, AGGREGATION SELECT only uses the final predicted answer and therefore remains an outcome-level selection strategy.

Formally, for the i-th instance, let $\mathcal{C}_i = \{C_i^{(k)}\}_{k=1}^{K}$ denote the K reasoning trajectories generated by the pretrained MLLM, where each trajectory ends with a predicted answer $\hat{y}_i^{(k)} \in \mathcal{O}_i$. We first compute the answer confidence score for each trajectory using the same margin-based definition as in the main text:

$$\Delta_{\text{ans}}(C_i^{(k)}) = \frac{1}{|\hat{y}_i^{(k)}|} \sum_{a_t \in \hat{y}_i^{(k)}} \left(p(a_t^{(1)} \mid a_{<t}) - p(a_t^{(2)} \mid a_{<t})\right) \quad (4)$$

where $a_t^{(1)}$ and $a_t^{(2)}$ denote the top-1 and top-2 token predictions at position t, respectively.

For each candidate answer $y \in \mathcal{O}_i$, we accumulate the confidence scores of all trajectories that predict y:

$$S_{\text{agg}}(y) = \sum_{k:\, \hat{y}_i^{(k)} = y} \Delta_{\text{ans}}(C_i^{(k)}). \quad (5)$$

The confidence-supported answer is selected as

$$\hat{y}_i^{\star} = \arg\max_{y \in \mathcal{O}_i} S_{\text{agg}}(y). \quad (6)$$

Finally, among all trajectories predicting $\hat{y}_i^{\star}$, we select the trajectory with the highest answer confidence as the pseudo-label:

$$C_i^{\text{agg}} = \arg\max_{C_i^{(k)}:\, \hat{y}_i^{(k)} = \hat{y}_i^{\star}} \Delta_{\text{ans}}(C_i^{(k)}). \quad (7)$$

Algorithm 2 AGGREGATION SELECT

Input: trajectories $\{C_i^{(k)}\}_{k=1}^{K}$ with final answers $\{\hat{y}_i^{(k)}\}_{k=1}^{K}$; **Output:** pseudo-label path C_i^{agg}

for $k = 1$ to K **do**
  Compute answer confidence $\Delta_{\text{ans}}(C_i^{(k)})$
end for
for each answer option $y \in \mathcal{O}_i$ **do**
  $S_{\text{agg}}(y) \leftarrow \sum_{k:\, \hat{y}_i^{(k)} = y} \Delta_{\text{ans}}(C_i^{(k)})$
end for
$\hat{y}_i^{\star} \leftarrow \arg\max_{y \in \mathcal{O}_i} S_{\text{agg}}(y)$
$C_i^{\text{agg}} \leftarrow \arg\max_{C_i^{(k)}:\, \hat{y}_i^{(k)} = \hat{y}_i^{\star}} \Delta_{\text{ans}}(C_i^{(k)})$
return C_i^{agg}

This strategy favors answers that receive strong collective confidence support while retaining the most confident trajectory within the winning answer group. However, because it does not inspect intermediate reasoning steps, it may still propagate trajectories that are answer-confident but logically flawed. The complete procedure is summarized in Algorithm 2.

B Details of Datasets

We conduct experiments on three multimodal reasoning benchmarks: M^{3}CoT, ScienceQA, and GeoQA. These datasets cover complementary reasoning scenarios, including multi-domain visual-textual reasoning, science question answering, and theorem-based geometry deduction. Since practical preprocessing may remove invalid, non-decodable, or format-incompatible samples, Table 4 reports only the official benchmark scale rather than the exact processed training count used in our pipeline. The processed split statistics are provided in the released code for full reproducibility.

Dataset	Scale	Format	Reasoning Focus
M^{3}CoT	11,459	MCQ	Multi-domain multimodal CoT
ScienceQA	21,208	MCQ	Science knowledge grounding and causal reasoning
GeoQA	4,998	MCQ	Geometry diagram understanding and theorem deduction

Table 4: Official benchmark statistics. "Scale" denotes the dataset scale reported by the original benchmark papers; the exact number of samples after preprocessing may differ slightly.

M^{3}CoT. M^{3}CoT (Chen et al., 2024) is a multi-modal Chain-of-Thought benchmark designed for multi-domain, multi-step, and multi-modal reasoning. It contains questions from science, mathematics, and commonsense domains, with diverse visual inputs such as diagrams, plots, infographics, and natural images. Each instance is formulated as a multiple-choice problem with visual context, textual input, answer options, and reasoning annotations. Compared with standard VQA-style benchmarks, M^{3}CoT places greater emphasis on compositional reasoning, cross-modal grounding, and step-wise inference.

Rationale. The red train cart is shaped like a small house, suggesting that it is used as a living space rather than for ordinary cargo transportation. The visible door further indicates that the cart is meant to be entered and occupied. Therefore, the correct answer is (B) "Providing living quarters."

Figure 5: Example from M^{3}CoT. Question: *What is the purpose of the red train cart?* Options: (A) Transporting cargo, (B) Providing living quarters, (C) Carrying passengers, (D) Storing equipment. **Answer: (B)**.

ScienceQA. ScienceQA (Lu et al., 2022) is a large-scale multimodal science question answering benchmark collected from elementary and high-school science curricula. It contains multiple-choice questions paired with multimodal contexts such as diagrams, charts, illustrations, and textual descriptions. In our setting, ScienceQA evaluates whether the model can ground scientific concepts in multimodal inputs and follow the underlying causal or relational structure of the problem.

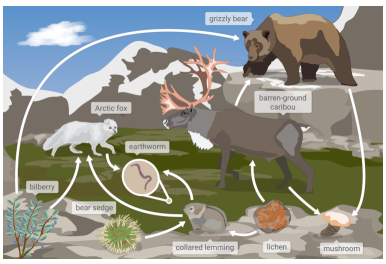

Figure 6: Example from ScienceQA. Question: *Which of these organisms contains matter that was once part of the lichen?* Options: (A) bilberry, (B) mushroom. **Answer: (B)**.

Rationale. The arrows in the food web indicate how matter flows between organisms. There is no arrow path from the lichen to the bilberry, so matter from the lichen does not flow to the bilberry. In contrast, the diagram supports a path from the lichen to the mushroom. Therefore, the correct answer is (B).

GeoQA. GeoQA (Chen et al., 2021) is a multimodal numerical reasoning benchmark for geometry problem solving. Each problem contains a textual question, an accompanying diagram, and four answer options. GeoQA emphasizes theorem-based deduction and diagram understanding, covering geometric concepts such as triangle properties, parallel lines, similarity, length relations, and angle reasoning. Since the original problems are written in Chinese, English translations are used only for illustrative examples in this paper.

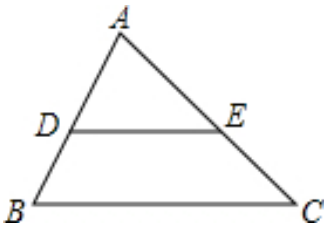

Figure 7: Example from GeoQA. Question: *In triangle ABC, given that* $\angle A = 80°$, $\angle B = 60°$, *and* $DE \parallel BC$, *what is the value of* $\angle CED$? Options: (A) 40°, (B) 60°, (C) 120°, (D) 140°. **Answer: (D)**.

Rationale. By the triangle angle-sum theorem, $\angle C = 180° - 80° - 60° = 40°$. Since $DE \parallel BC$, $\angle C$ and $\angle CED$ form a pair of same-side interior angles and are supplementary. Thus, $\angle CED = 180° - 40° = 140°$, and the correct answer is (D).

Splits and label usage. For all datasets, we follow the official or released preprocessing splits used by prior work whenever applicable. In the answer-free setting, ground-truth answers and human-written rationales are not used for pseudo-label construction or preference optimization. They are used only for final evaluation and for supervised reference baselines.

Prompt Instruction for Reasoning Evaluation

You are a logical reasoning expert. Your task is to evaluate 5 reasoning paths generated for the same multimodal multiple-choice question, and identify the single best one that is logically rigorous and whose final option is best supported by its own reasoning and the given answer options.
These paths were generated from questions involving images, but you can only see the text. You must judge purely based on the reasoning quality and content logic.
Please ignore any response that directly gives a final option (e.g., "So, the answer is B.") without any explanation.
You should evaluate from the following aspects:

- Is the reasoning logically consistent, with no contradictions or jumps?
- Is the final option well supported by the reasoning steps?
- Is the reasoning focused on the question and answer options?
- Is the expression clear, well-structured, and easy to understand?
- Is the final option consistent with the reasoning and the given options?

Figure 8: Prompt used for Round-0 text-only expert selection. The expert is asked to select the most logically consistent reasoning path based only on the question, answer options, and candidate textual trajectories.

C Prompt Templates for Expert Discussion

To facilitate structured text-only expert collaboration, we use standardized prompt templates for both Round-0 selection and Round-1 reflective discussion. In Round 0, each expert independently receives the original question, answer options, and the top candidate reasoning trajectories selected by confidence filtering. The image is intentionally

excluded, forcing experts to evaluate only textual coherence, deductive validity, and whether the final option is supported by the candidate rationale itself. In Round 1, when no majority agreement is reached, each expert reviews the initial selections and rationales of the other experts, then decides whether to keep or revise its original choice.

Prompt for Agent A's Reflective Evaluation

You are Agent A. You and Agent B, Agent C have been assigned the same task:

Task Instruction: You are a logical reasoning expert. Your task is to evaluate 5 reasoning paths generated for the same multimodal multiple-choice question, and identify the single best one that is logically rigorous and whose final option is best supported by its own reasoning and the given answer options.
These paths were generated from questions involving images, but you can only see the text. You must judge purely based on the reasoning quality and content logic.
Please ignore any response that directly gives a final option (e.g., "So, the answer is B.") without any explanation.

Evaluation Criteria:

- Is the reasoning logically consistent, with no contradictions or jumps?
- Is the final option well supported by the reasoning steps?
- Is the reasoning focused on the question and answer options?
- Is the expression clear, well-structured, and easy to understand?
- Is the final option consistent with the reasoning and the given options?

Context: You initially selected path `<A_index>` with final option "`<A_answer>`".
Your rationale: `<A_rationale>`
Agent B selected path `<B_index>` with final option "`<B_answer>`" and rationale: `<B_rationale>`
Agent C selected path `<C_index>` with final option "`<C_answer>`" and rationale: `<C_rationale>`

Now, after reviewing both Agent B and Agent C's reasoning, please reflect and respond:

- Re-evaluate whether the path you originally selected is still the most reasonable based on all information.
- Consider whether Agent B or Agent C's path offers stronger logic or better alignment with the evaluation criteria.
- If you change your selection, explain why. If you keep your choice, explain why it still stands.

Please respond strictly in the following format:

Best Path: <0-4>
Rationale: <your updated explanation >
Judgment on B's reasoning: <agree/disagree and why >
Judgment on C's reasoning: <agree/disagree and why >

Figure 9: Prompt used for Round-1 reflective discussion. The expert reconsiders its initial selection after reviewing the choices and rationales of the other experts.

Expert-call cost. Each instance requires three Round-0 text-only expert calls. Round-1 deliberation is invoked only when the three experts fail to reach a majority agreement, adding at most three additional calls. In practice, fewer than 0.1% of samples remain unresolved after deliberation. Since only the top-$M = 5$ confidence-filtered trajectories are sent to the experts, the verification cost is bounded and does not grow with the total number of initially sampled paths K.

We do not report standard deviations or significance tests in the main tables, following common practice in related answer-free and self-improvement studies (Wei et al., 2025; Deng et al., 2024; Zhao et al., 2025). The observed improvements are consistent across all three datasets and across both outcome accuracy and process-level PRM scores.

Item	Configuration
Backbone	Qwen2.5-VL-7B-Instruct
Context length	8k
Training framework	LLaMA-Factory
Hardware	8×A100 80GB GPUs
Fine-tuning type	Full-parameter fine-tuning
Precision	`bf16`
Memory optimization	FlashAttention-V2, DeepSpeed ZeRO-3
SFT epochs	2
DPO epochs	1
Batch size	8
Learning rate	5×10^{-7}
Scheduler	Cosine decay with 0.1 warm-up ratio
DPO β	0.2
Auxiliary SFT loss weight μ	1.0
Generated paths per sample	$K = 10$
Retained candidate paths	$M = 5$
Expert committee size	$E = 3$
Expert models	Qwen-Max, QwQ, DeepSeek-R1

Table 5: Implementation details for trajectory generation, expert verification, SFT, and DPO.

D Analysis of GPT-4o Generation and Distillation

Strategy	CoT	M^{3}CoT	ScienceQA	GeoQA	Avg.
GPT-4o Prompted Inference					
GPT-4o Direct[a]	×	60.15	84.77	53.02	65.98
GPT-4o CoT[a]	✓	42.32	66.90	32.07	47.10
GPT-4o Tuned CoT[a]	✓	43.74	69.67	54.02	55.81

[a]Uses top-p sampling instead of greedy decoding.

Table 6: Zero-shot GPT-4o prompted inference results. Only Greedy-style single-response accuracy is reported because GPT-4o is not evaluated with our multi-path aggregation protocol.

Table 6 and Table 7 compare MAGNET with GPT-4o under different prompting and distillation strategies. We split the original wide table into two single-column tables so that the results can be placed closer to the corresponding discussion rather than floating to the top of a later page.

Strategy	M^{3}CoT		ScienceQA		GeoQA		Avg.	
	Greedy	Agg.	Greedy	Agg.	Greedy	Agg.	Greedy	Agg.
SFT on GPT-4o Generated Outputs								
SFT on Direct	62.35	63.53	79.45	86.00	57.81	63.26	66.54	70.93
SFT on GPT-4o CoT	28.33	44.35	12.85	66.83	11.33	29.81	17.50	47.00
SFT on Tuned CoT	22.73	51.08	24.62	78.05	20.60	20.03	22.65	49.72
Answer-Free Post-Training								
Ours	**67.17**	**70.10**	**86.08**	**87.93**	**72.28**	**83.82**	**75.18**	**80.62**

Table 7: Accuracy comparison between GPT-4o distillation and MAGNET. GPT-4o-generated outputs are treated as pseudo-supervision rather than ground-truth labels.

In the zero-shot setting, GPT-4o is evaluated

with three prompt variants. *Direct Answer* asks the model to output only the answer. *CoT Prompt* elicits step-by-step rationales using a standard reasoning instruction. *Tuned CoT Prompt* is manually optimized to better fit GPT-4o's response style. Although CoT prompting is often expected to improve reasoning transparency, both standard and tuned CoT prompts substantially reduce GPT-4o's accuracy compared with direct answering.

We further examine a distillation setting where smaller MLLMs are fine-tuned on GPT-4o outputs generated under each prompt type. The resulting models underperform GPT-4o itself and, in some cases, even fall below zero-shot baselines. This supports the small-model learnability gap (Li et al., 2025): reasoning traces generated by a stronger teacher may contain complex cross-modal assumptions that are difficult for a smaller student model to imitate. In contrast, MAGNET avoids external teacher distillation and constructs preference signals from the target model's own trajectories, making the supervision more capacity-aligned and more transferable.

E Analysis of Discarded Cases in Round 1

To examine the effect of difficult or ambiguous samples that are discarded after multi-agent deliberation, we conduct an additional ablation study. For samples where no consensus is reached in either Round 0 or Round 1, we use DeepSeek-R1 as a strong single-agent adjudicator to select a final trajectory in Round 2. These additional Round-2 selections are then incorporated into SFT or DPO training, allowing the model to reuse samples that are otherwise discarded.

Strategy	M^3CoT		ScienceQA		GeoQA		Avg.	
	Greedy	Agg.	Greedy	Agg.	Greedy	Agg.	Greedy	Agg.
Recovered non-consensus cases								
SFT Round-2	67.00	69.28	83.33	85.78	70.95	80.11	73.76	78.39
DPO Round-2	65.96	69.93	83.77	86.49	71.35	82.76	73.69	79.73
Final setting								
Ours	**67.17**	**70.10**	**86.08**	**87.93**	**72.28**	**83.82**	**75.18**	**80.62**

Table 8: Accuracy results when incorporating Round-2 samples selected by DeepSeek-R1 from previously non-consensus cases.

Tables 8 and 9 show that including these ambiguous cases does not consistently improve performance. Although Round-2 training recovers more pseudo-labeled data, it introduces noisier supervision and slightly underperforms our final method, especially in PRM scores. This suggests that non-consensus samples tend to be inherently uncertain or logically ambiguous. Forcing a decision on them, even with a strong adjudicator, can introduce biased or unreliable preference signals. Therefore, our final framework excludes non-consensus cases to preserve pseudo-label quality and training stability.

Strategy	M^3CoT		ScienceQA		GeoQA		Avg.	
	Greedy	Agg.	Greedy	Agg.	Greedy	Agg.	Greedy	Agg.
Recovered non-consensus cases								
SFT Round-2	90.56	91.24	93.95	94.49	80.56	81.10	88.36	88.94
DPO Round-2	90.94	91.25	93.65	94.81	80.66	81.56	88.42	89.21
Final setting								
Ours	**91.53**	**91.42**	**94.60**	**94.97**	**81.68**	**82.80**	**89.27**	**89.73**

Table 9: PRM score comparison when incorporating Round-2 samples selected by DeepSeek-R1. The slight decline indicates that ambiguous non-consensus cases introduce lower-quality reasoning supervision.

F DPO Loss Formulation

The standard DPO loss (Rafailov et al., 2023) aligns the policy model π_θ with preferred trajectories relative to a frozen reference model π_{ref}:

$$\begin{aligned}\mathcal{L}_{\text{DPO}}(\theta) &= -\mathbb{E}_{(x_i, y_i^+, y_i^-)\sim\mathcal{D}_{\text{pref}}}\left[\log\sigma\left(\beta(\Delta f_\theta(x_i) - \Delta f_{\text{ref}}(x_i))\right)\right], \\ \Delta f_s(x_i) &= f_s(x_i, y_i^+) - f_s(x_i, y_i^-), \quad s \in \{\theta, \text{ref}\},\end{aligned} \tag{8}$$

where f_θ denotes the model's log-probability scoring function, and β controls preference discrimination sharpness.

G Tabulated Values for Stage-Wise Figures

To provide precise numeric references for the stage-wise plots in the main paper, we report the corresponding accuracy and PRM values in Tables 10 and 11. These tables show how performance changes from the zero-shot baseline to confidence-filtered SFT stages and final DPO alignment.

Strategy	M^3CoT		ScienceQA		GeoQA		Avg.	
	Greedy	Agg.	Greedy	Agg.	Greedy	Agg.	Greedy	Agg.
Zero-shot CoT	57.81	67.28	76.14	84.65	46.02	70.56	59.99	74.16
SFT Round-0	62.25	67.77	82.63	85.62	71.49	81.17	72.12	78.19
SFT Round-1	**67.69**	69.15	85.57	87.89	**73.74**	82.23	**75.67**	79.76
Ours	67.17	**70.10**	**86.08**	**87.93**	72.28	**83.82**	75.18	**80.62**

Table 10: Accuracy values corresponding to the stage-wise accuracy figure in the main text.

Strategy	M^3CoT		ScienceQA		GeoQA		Avg.	
	Greedy	Agg.	Greedy	Agg.	Greedy	Agg.	Greedy	Agg.
Zero-shot CoT	62.55	63.11	72.32	72.97	56.95	57.38	63.94	64.49
SFT Round-0	89.72	89.73	93.17	93.33	79.49	79.91	87.46	87.66
SFT Round-1	91.38	91.38	**95.00**	95.19	**81.86**	82.41	**89.41**	89.66
Ours	**91.53**	**91.42**	94.60	**94.97**	81.68	**82.80**	89.27	**89.73**

Table 11: PRM score values corresponding to the stage-wise PRM figure in the main text.