FROM STAGES TO STATES: UNIFIED ITERATIVE REASONING FOR MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Chain-of-Thought (MCoT) substantially enhances the complex reasoning capabilities of Multimodal Large Language Models (MLLMs) by explicitly modeling intermediate reasoning processes. However, most learning-based MCoT methods follow a stage-centric paradigm, where different reasoning behaviors are assigned to predefined stages and optimized separately. This results in fragmented reasoning and a training-inference reasoning-state gap. To address these limitations, we propose ***ReState***, a unified iterative reasoning framework that reformulates multimodal reasoning from a stage-driven pipeline into a state-driven dynamic process. It comprises two complementary components: Unified Reasoning State Modeling (URSM) and Closed-Loop Reasoning-State Learning (CRSL). URSM unifies reasoning construction, correct reasoning preservation, and erroneous reasoning correction as reasoning-state transitions within a shared state space, enabling these heterogeneous reasoning behaviors to be learned jointly. CRSL further incorporates self-generated historical reasoning states during training and recursively updates the current state during inference, thereby mitigating the reasoning-state gap and enabling iterative self-refinement without requiring an additional reflection module, critic, or independent correction stage. Furthermore, we construct VersaCoT-10K, a high-quality CoT dataset spanning eight visual reasoning domains, which provides diverse supervision for multi-domain reasoning and reasoning-state refinement. Extensive experiments demonstrate that ReState exhibits strong adaptability across different MLLM backbones and achieves strong performance on multiple public reasoning benchmarks, supporting the shift from stage-centric reasoning toward state-centric reasoning. Our data and code will be open source.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have achieved remarkable progress across a wide range of tasks, including visual question answering (Yu et al., 2026; Cheng et al., 2025; Yang et al., 2025b; Chen et al., 2024), scientific reasoning (Lu et al., 2022; Wang et al., 2024a; Li et al., 2024; Burgess et al., 2025; Zhang et al., 2025), and chart understanding (Masry et al., 2022; Xu et al., 2025; Masry et al., 2025; Yang et al., 2025a). As their perceptual and cross-modal capabilities improve, research has increasingly shifted from direct answer prediction toward explicit modeling of intermediate reasoning processes (Zhang et al., 2024; Cheng et al., 2026; Dong et al., 2026). Multimodal Chain-of-Thought (MCoT) follows this direction by introducing explicit intermediate reasoning to decompose complex problems into interconnected reasoning steps (Wang et al., 2025).

Existing MCoT research mainly focuses on constructing and optimizing reasoning chains. Multimodal-CoT (Zhang et al., 2024) introduced a two-stage paradigm of CoT generation and answer inference, while subsequent methods incorporated structured reasoning representations (Mitra et al., 2024), visual grounding (Shao et al., 2024), and other mechanisms (Tan et al., 2024a; Wang et al., 2024a; He et al., 2024). More recently, MIND (Yu et al., 2026) jointly learns positive and negative CoTs, extending MLLMs from reasoning generation toward reasoning discrimination and correction. Despite these advances, most learning-based MCoT methods remain *stage-centric*: different reasoning behaviors are assigned to predefined stages with stage-specific models, mappings, or objectives. This leads to two limitations. First, reasoning behaviors are modeled separately, resulting in fragmented reasoning without a unified mechanism for state evolution. Second, downstream

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

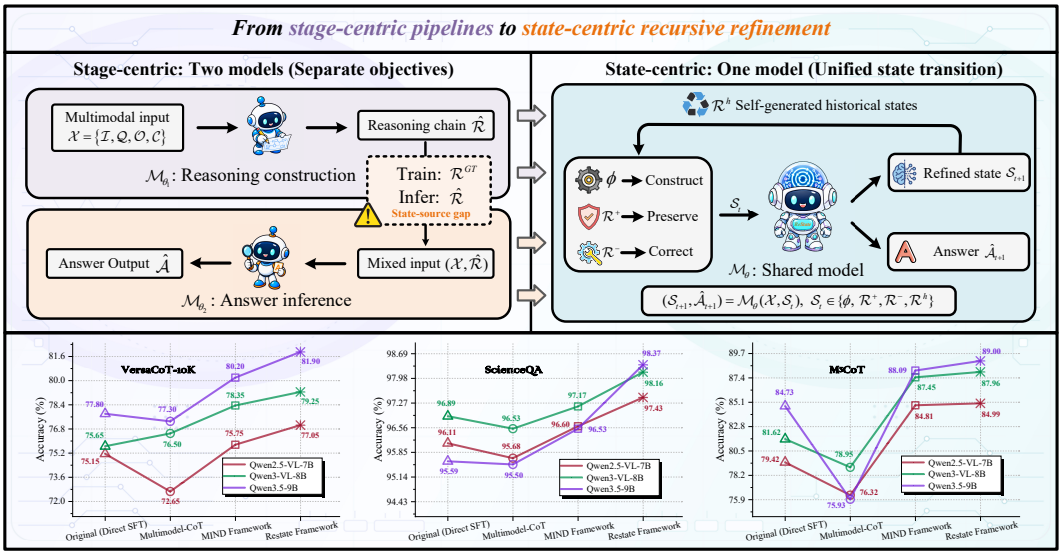


Figure 1: Comparison of reasoning paradigms and performance. The stage-centric paradigm assigns different reasoning behaviors to predefined stages, leading to fragmented modeling and a training-inference reasoning-state gap. In contrast, ReState proposes a state-centric paradigm that recursively refines explicit reasoning states with a unified state-update operator, unifying reasoning construction, preservation, and correction while leveraging model-generated historical states for closed-loop learning. Across three backbones, ReState improves performance by **0.90%-4.60%**, **0.83%-2.87%**, and **0.18%-13.07%** on VersaCoT-10K, ScienceQA, and M^{3}CoT, respectively.

stages are typically trained on annotated or pre-constructed reasoning states but receive upstream model-generated states during inference, introducing a training-inference reasoning-state gap.

Rather than designing increasingly complex reasoning stages, we ask whether heterogeneous reasoning behaviors can instead be unified as the evolution of explicit reasoning states. Motivated by this view, we propose ***ReState*** (Reasoning-State Refinement), a unified iterative reasoning framework that reformulates multimodal reasoning from a stage-driven pipeline into a state-driven dynamic process. As illustrated in Fig. 1, ReState focuses not on which predefined stage the model is in, but on what reasoning state it currently has and how that state should be updated. Reasoning states of different sources and qualities can therefore be handled by a shared state-update operator.

Specifically, ReState consists of two complementary designs. First, Unified Reasoning State Modeling (URSM) reformulates reasoning construction, correct reasoning preservation, and erroneous reasoning correction as transitions within a shared state space. Given an empty, correct, or erroneous reasoning state, the same MLLM learns to update it toward a correct reasoning state, unifying previously fragmented reasoning behaviors within a common state-evolution process. Second, Closed-Loop Reasoning-State Learning (CRSL) incorporates self-generated historical reasoning states during training and recursively feeds the current state back into the same model during inference. This closed loop of "state generation $\rightarrow$ state feedback $\rightarrow$ state update" reduces the training-inference reasoning-state gap and enables iterative self-refinement without additional reflection modules, critics, or independent correction stages. From Fig. 1, ReState consistently improves performance across multiple datasets and backbones, with gains ranging from 0.18% to 13.07%. To further support multi-domain reasoning and reasoning-state refinement, we construct VersaCoT-10K, a high-quality CoT dataset containing 10,000 single-image multiple-choice samples across eight heterogeneous visual reasoning domains for research. The main contributions are summarized as follows:

- We propose ReState, a unified iterative reasoning framework for MLLMs. By reformulating the reasoning process from a stage-driven pipeline paradigm to a state-driven dynamic process, ReState enables unified organization and continuous refinement of heterogeneous reasoning behaviors while establishing a paradigm shift from stage-centric to state-centric reasoning.
- To address the fragmented modeling of reasoning behaviors in stage-based methods, we propose the URSM mechanism, which unifies heterogeneous reasoning behaviors as reasoning-state transitions,

enabling joint optimization of reasoning construction, correct reasoning preservation, and erroneous reasoning correction within a unified state space.

- To bridge the reasoning-state gap between training and inference, we propose the CRSL strategy, which learns from self-generated reasoning states during training and recursively refines them during inference, enabling continuous reasoning-state refinement without introducing additional reflection modules or independent correction stages.
- We construct VersaCoT-10K, a high-quality CoT dataset covering eight heterogeneous visual reasoning scenarios, providing new training and evaluation resources for studying multi-domain reasoning and reasoning-state refinement in MLLMs.

2 Related Work

Prompt-based MCoT Reasoning elicits explicit reasoning from MLLMs through prompt design, in-context examples, or external information without parameter updates (Wang et al., 2025). Early approaches mainly relied on few-shot demonstrations and multi-path sampling (Wei et al., 2022; Wang et al., 2022), while later studies strengthened visual-language interaction through explicit visual descriptions, structured representations, and richer intermediate reasoning forms (Chen et al., 2023; Mitra et al., 2024; Meng et al., 2023; Hu et al., 2024; Cheng et al., 2026). Retrieval and tool use further extend reasoning with external knowledge and specialized visual capabilities (Lu et al., 2023; You et al., 2023; Wu et al., 2024b). Despite their flexibility, these methods remain sensitive to model capability, prompt design, and sampling strategy (Jin et al., 2026).

Learning-based MCoT Reasoning explicitly supervises intermediate reasoning to connect visual evidence, reasoning processes, and final answers. Multimodal-CoT (Zhang et al., 2024) first established a two-stage paradigm of CoT generation and answer inference. Building on this, MC-CoT (Tan et al., 2024a), DPMM-CoT (He et al., 2024), DDCoT (Zheng et al., 2023), and T-SciQ (Wang et al., 2024a) enhance reasoning learning from the perspectives of multi-path collaboration, cross-modal semantic alignment, responsibility decomposition, and knowledge transfer, respectively. Recent studies have further extended MCoT through directions such as visual grounding, interleaved representations, long-chain reasoning, and reinforcement learning (Li et al., 2025; Man et al., 2025; Chen et al., 2026; Wu et al., 2024a; Lee et al., 2024; Dong et al., 2025; Tan et al., 2026; Huang et al., 2026). In addition, MIND (Yu et al., 2026) further learns from positive and negative CoTs, extending MLLMs from reasoning generation to reasoning discrimination and correction. Nevertheless, existing methods generally treat intermediate CoTs as stage-specific outputs rather than reusable states for recursive refinement. This stage-centric design results in fragmented reasoning and a training–inference reasoning-state gap, motivating our shift toward state-centric reasoning.

3 Preliminary: Stage-centric Reasoning Paradigm

Most existing learning-based MCoT methods follow a stage-centric reasoning paradigm (Zhang et al., 2024; Tan et al., 2024a; He et al., 2024; Zheng et al., 2023; Wang et al., 2024a; Yu et al., 2026), typically decomposing reasoning into CoT generation and answer inference. These stages are modeled independently rather than through unified reasoning-state evolution.

In the CoT generation stage, the model generates an explicit reasoning chain from the input:

$$\hat{\mathcal{R}} = \mathcal{M}_{\theta_1}(\mathcal{X}), \quad \mathcal{X} = \{\mathcal{I},\ \mathcal{Q},\ \mathcal{O},\ \mathcal{C}\} \tag{1}$$

where $\mathcal{I}$, $\mathcal{Q}$, $\mathcal{O}$ and $\mathcal{C}$ denote the image, question, candidate options, and image caption, respectively. $\mathcal{M}_{\theta_1}$ denotes an MLLM parameterized by θ_1. This stage is supervised using annotated reasoning chains $\hat{\mathcal{R}}^{GT}$, learning a mapping from multimodal inputs to explicit reasoning processes.

In the answer inference stage, the generated reasoning chain is used as an additional condition:

$$\hat{\mathcal{A}} = \mathcal{M}_{\theta_2}(\mathcal{X}, \hat{\mathcal{R}}) \tag{2}$$

where $\hat{\mathcal{A}}$ denotes the predicted answer, $\mathcal{M}_{\theta_2}$ denotes an MLLM parameterized by θ_2. Notably, this stage is trained on annotated reasoning $\mathcal{R}^{GT}$ but receives model-generated reasoning $\hat{\mathcal{R}}$ by the preceding stage at inference, resulting in a training-inference reasoning-state gap.

Thus, predefined reasoning stages lead to fragmented reasoning and a training-inference reasoning-state gap. These limitations motivate a unified formulation based on reasoning-state evolution.

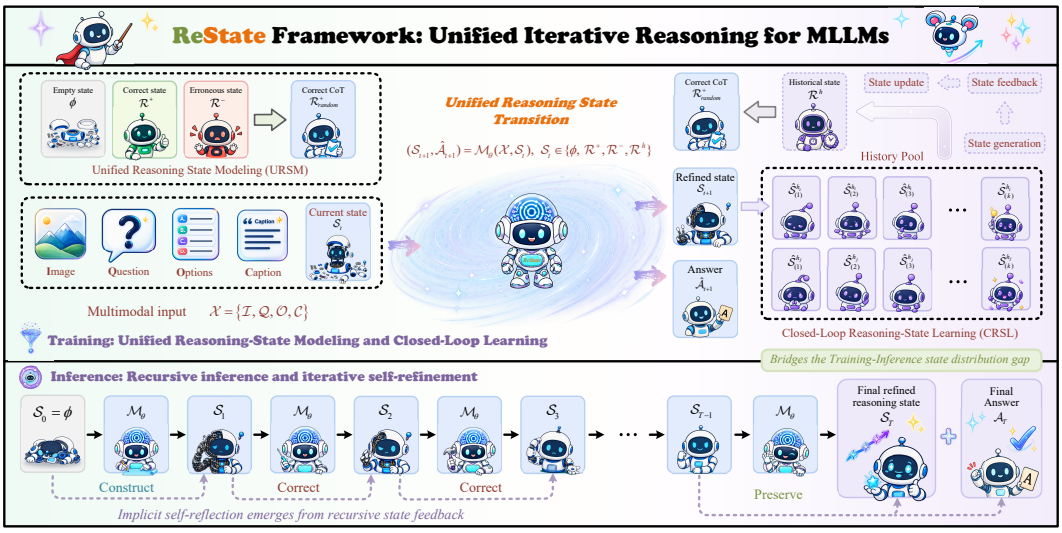


Figure 2: Overview of the ReState framework. During training, URSM learns a shared state-update operator from empty, correct, and erroneous reasoning states, while CRSL feeds model-generated historical states back into training to reduce the training-inference reasoning-state mismatch. At inference, the same model recursively updates the current reasoning state and answer, enabling iterative self-refinement without additional critics, reflection modules, or separate correction stages.

4 ReState: A Unified Iterative Reasoning Framework

4.1 Overview

To address the limitations of stage-centric reasoning in Section 3, we propose ReState, which reformulates multimodal reasoning from a stage-driven process into a unified dynamic evolution of explicit reasoning states. As shown in Fig. 2, ReState consists of unified state learning during training and recursive state refinement during inference, built upon two complementary designs: URSM and CRSL. During training, unlike existing MCoT methods (Yu et al., 2026), URSM unifies reasoning construction, correct reasoning preservation, and erroneous reasoning correction as reasoning-state transitions implemented by the same MLLM. Building on URSM, CRSL further incorporates self-generated historical states to expose the state-update operator to model-induced states and reduce the training-inference reasoning-state gap. During inference, ReState starts from an empty state and recursively feeds the current reasoning state back into the same state-update operator, progressively preserving valid reasoning, correcting potential errors, and updating the answer. The entire process reuses the same model throughout and achieves iterative self-refinement without requiring an additional reflection module, critic, or independent correction stage.

4.2 Unified Reasoning State Modeling (URSM) Mechanism

Existing MCoT methods typically treat reasoning chains as stage-specific intermediate outputs. In contrast, URSM represents intermediate reasoning of different qualities as continuously updatable states and models them with a shared state-update operator.

$$(\mathcal{S}_{t+1}, \hat{\mathcal{A}}_{t+1}) = \mathcal{M}_\theta(\mathcal{X}, \mathcal{S}_t), \quad \mathcal{S}_t \in \{\phi,\ \mathcal{R}^+, \mathcal{R}^-\} \tag{3}$$

where $\mathcal{M}_\theta$ is the shared reasoning-state transition function implemented by a single MLLM parameterized by θ. $\mathcal{S}_t$, $\mathcal{S}_{t+1}$ and $\hat{\mathcal{A}}_{t+1}$ denote the current reasoning state, updated reasoning state and predicted answer. ϕ, $\mathcal{R}^+, \mathcal{R}^-$ denote the empty, valid, and erroneous reasoning states, respectively, with ϕ represented by the special token $[EMPTY_SOLUTION]$.

Within this unified state space, heterogeneous reasoning behaviors are reformulated as different reasoning-state transitions. When $\mathcal{S}_t = \phi$, the model constructs an initial reasoning state from the multimodal input, corresponding to reasoning construction. When $\mathcal{S}_t = \mathcal{R}^+$, it preserves valid reasoning evidence while improving completeness and consistency, corresponding to correct

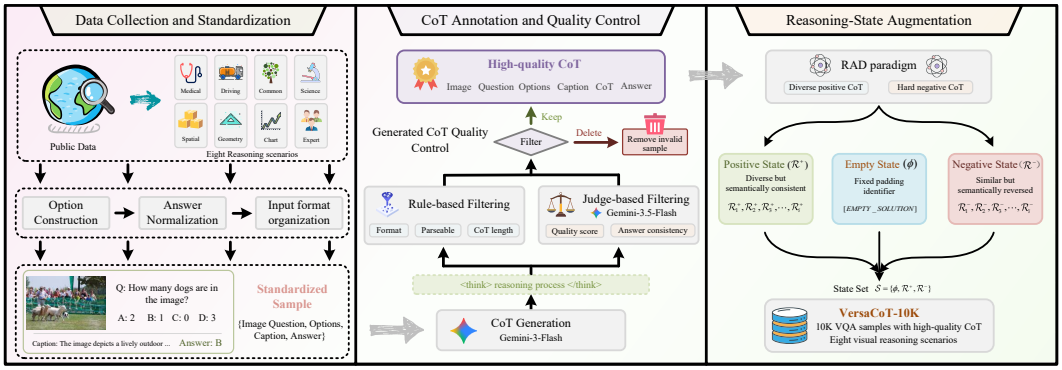


Figure 3: Construction pipeline of VersaCoT-10K dataset. It consists of three stages: *Data Collection and Standardization*, *CoT Annotation and Quality Control*, and *Reasoning-State Augmentation*.

reasoning preservation. When $\mathcal{S}_t = \mathcal{R}^-$, it revisits the original multimodal evidence to correct errors in the current state, corresponding to erroneous reasoning correction. By jointly learning these transitions, the model moves beyond generating reasoning chains from scratch to understanding, re-evaluating, and refining existing reasoning states, extending its capability from reasoning generation to reasoning-state understanding and refinement.

4.3 Closed-loop Reasoning-State Learning (CRSL) Strategy

Although URSM provides a unified reasoning-state transition mechanism, training only on pre-constructed states cannot fully cover the model-generated states encountered during recursive inference. Differences in their source and error patterns therefore introduce a training–inference reasoning-state gap, limiting the model's ability to refine its own historical states.

To address this issue, CRSL feeds self-generated historical reasoning states back into training, extending the URSM state space to $\mathcal{S}_t \in \{\phi,\ \mathcal{R}^+, \mathcal{R}^-, \mathcal{R}^h\}$. For each training sample i, CRSL maintains a sample-specific, fixed-capacity history pool:

$$\mathcal{H}_i = \{\hat{\mathcal{S}}^i_{(1)}, \hat{\mathcal{S}}^i_{(2)}, \ldots, \hat{\mathcal{S}}^i_{(k)}\} \tag{4}$$

where $\hat{\mathcal{S}}^i_{(k)}$ denotes the reasoning state generated for sample i at the k_{th} round. The pool is updated in a first-in-first-out manner, allowing $\mathcal{H}_i$ to continuously track the state distribution induced by the model throughout training. In our experiments, the pool capacity is set to 2.

By reusing these historical states as subsequent training inputs, CRSL forms a closed loop of "state generation $\rightarrow$ state feedback $\rightarrow$ state update". This exposes the state-update operator to model-induced states closer to those encountered during recursive inference, thereby reducing the reasoning-state gap. Moreover, revisiting historical states encourages the model to re-evaluate and correct its reasoning, providing a basis for iterative self-refinement.

At inference, ReState recursively applies the same state-update operator from an empty state:

$$\mathcal{S}_0 = \phi, \quad (\mathcal{S}_{t+1}, \hat{\mathcal{A}}^{t+1}) = \mathcal{M}_\theta(\mathcal{X}, \mathcal{S}_t), \quad t = 0, \ldots, \mathcal{T}-1 \tag{5}$$

This produces the state-evolution trajectory $\mathcal{S}_0 \rightarrow \mathcal{S}_1 \rightarrow \cdots \rightarrow \mathcal{S}_T$, $\hat{\mathcal{A}} = \hat{\mathcal{A}}^T$. At each iteration, the model feeds back the current state to preserve valid reasoning and correct potential errors. Because the entire recursive process reuses the same state-update operator, ReState achieves iterative self-refinement without additional reflection modules, critics, or independent correction stages.

5 VersaCoT-10K Dataset

Existing multimodal reasoning datasets often lack explicit CoT annotations, broad domain coverage, or diverse reasoning states for reasoning refinement. To address these limitations, we construct VersaCoT-10K, a high-quality dataset containing 10,000 single-image multiple-choice samples across eight heterogeneous visual reasoning domains. As shown in Fig. 3, VersaCoT-10K is constructed

Table 1: Main results (%) on the VersaCoT-10K dataset. MED = medical, DRIVE = driving and safety, COMMON = visual commonsense, DIAG = scientific diagram, SPAT = spatial perception, GEO = geometric and mathematical, CHART = chart computation, and MULTI = multidisciplinary knowledge. Red denotes the best result, and blue denotes the second best result.

Model	Learning	Size	MED	DRIVE	COMMON	DIAG	SPAT	GEO	CHART	MULTI	**Avg**
Random	-	-	22.00	46.80	26.80	26.27	39.87	22.93	22.00	25.60	29.03
Human	-	-	66.00	89.20	90.40	88.00	92.00	94.40	96.00	85.20	87.65
Open-source Models + Zero-shot											
Qwen2.5-VL-7B	*Zero-shot*	7B	48.00	52.00	88.80	72.40	76.80	70.40	82.40	48.80	67.45
Qwen3-VL-8B	*Zero-shot*	8B	51.60	57.20	88.80	79.20	89.20	68.80	83.60	54.00	71.55
Qwen3.5-9B	*Zero-shot*	9B	61.60	59.60	87.60	81.20	83.60	70.80	84.80	54.00	72.90
Qwen3.5-122B-A10B	*Zero-shot*	122B	66.40	62.80	89.20	90.80	87.60	86.80	93.20	63.20	80.00
Qwen3.6-27B	*Zero-shot*	27B	68.80	59.60	90.00	88.40	86.40	87.60	93.60	62.00	79.55
MiniMax-M3	*Zero-shot*	428B	60.80	58.00	88.00	80.80	79.20	73.60	92.00	61.60	74.25
GLM-4.6V	*Zero-shot*	106B	60.40	59.20	90.00	85.20	87.60	78.40	93.60	60.80	76.90
Kimi-K2.6	*Zero-shot*	1T	68.40	63.20	91.60	77.20	85.20	55.20	94.00	58.00	74.10
Kimi-K3	*Zero-shot*	2.8T	71.60	70.40	90.00	92.40	90.80	91.60	98.00	80.40	85.65
Close-source Models + Zero-shot											
Qwen3.7-Flash	*Zero-shot*	-	65.60	62.80	84.00	86.80	84.80	74.40	90.00	60.00	76.05
Qwen3.8-Max	*Zero-shot*	-	74.00	70.00	92.40	91.60	92.00	94.80	96.80	71.20	85.35
Grok-4.6	*Zero-shot*	-	74.00	71.60	88.00	92.80	87.20	95.20	98.00	88.00	86.85
GLM-5V-Turbo	*Zero-shot*	-	66.00	63.60	89.20	86.00	91.20	64.00	94.80	59.60	76.80
GPT-5.6-Terra	*Zero-shot*	-	66.00	62.40	88.80	86.80	84.80	73.60	95.20	67.60	78.15
GPT-5.6-Sol	*Zero-shot*	-	70.80	69.20	90.00	92.00	82.80	87.60	97.20	74.40	83.00
Claude Sonnet 5	*Zero-shot*	-	68.80	63.60	89.20	86.80	82.00	78.80	94.80	69.20	79.15
Claude Opus 5	*Zero-shot*	-	70.80	75.20	89.20	90.80	85.60	89.60	98.40	78.00	84.70
Seed-1.8	*Zero-shot*	-	64.80	58.80	90.80	88.00	85.60	91.60	94.80	66.40	80.10
Seed-2.1-Pro	*Zero-shot*	-	74.40	71.20	89.60	90.40	88.80	96.80	94.80	75.60	85.20
Fine-tuning Methods											
Qwen2.5-VL-7B	*Fine-tune*	7B	**61.20**	65.60	**90.00**	83.20	88.40	72.40	91.20	49.20	75.15
Qwen2.5-VL-7B + Multimodal-CoT	*Fine-tune*	7B	49.20	62.00	83.60	82.40	81.20	73.20	92.40	**57.20**	72.65
Qwen2.5-VL-7B + MIND	*Fine-tune*	7B	55.20	66.00	88.80	85.20	85.60	74.00	**94.80**	56.40	75.75
Qwen2.5-VL-7B + ReState (Ours)	*Fine-tune*	7B	57.20	**68.00**	88.40	**87.60**	**89.60**	**76.80**	94.00	54.80	**77.05**
Qwen3-VL-8B	*Fine-tune*	8B	**59.20**	60.80	88.40	87.60	90.40	76.00	90.80	52.00	75.65
Qwen3-VL-8B + Multimodal-CoT	*Fine-tune*	8B	50.00	66.00	83.20	82.00	90.80	83.20	**95.60**	**61.20**	76.50
Qwen3-VL-8B + MIND	*Fine-tune*	8B	54.80	67.60	**90.80**	84.80	90.00	**83.60**	94.40	60.80	78.35
Qwen3-VL-8B + ReState (Ours)	*Fine-tune*	8B	58.80	**69.20**	87.20	**89.60**	**91.60**	82.80	94.40	60.40	**79.25**
Qwen3.5-9B	*Fine-tune*	9B	61.20	70.00	86.80	90.00	90.00	74.00	94.40	56.00	77.80
Qwen3.5-9B + Multimodal-CoT	*Fine-tune*	9B	56.40	63.60	83.60	87.20	88.00	**88.80**	94.00	56.80	77.30
Qwen3.5-9B + MIND	*Fine-tune*	9B	63.20	68.80	86.40	**91.60**	90.80	84.00	**96.00**	60.80	80.20
Qwen3.5-9B + ReState (Ours)	*Fine-tune*	9B	**66.40**	**76.00**	**89.20**	91.20	**92.00**	83.60	95.20	**61.60**	**81.90**

through three stages. First, we collect reasoning-intensive samples from multiple public benchmarks and standardize their options, answers, and multimodal inputs. Second, Gemini-3-Flash (Google DeepMind, 2025) generates visually grounded CoTs, followed by rule-based filtering and multimodal quality assessment with Gemini-3.5-Flash (Google DeepMind, 2026). Third, following RAD (Yu et al., 2026), each quality-filtered CoT is augmented with positive and hard negative reasoning states. Together with the empty state, they provide supervision for learning reasoning construction, correct reasoning preservation, and erroneous reasoning correction. Notably, historical states used by CRSL are instead generated dynamically during ReState training. Detailed construction procedures, prompts, quality-control criteria, and dataset statistics are provided in *Appendix* A and B.

6 Experiments

6.1 Experimental Setup

1) Datasets. We evaluate ReState on VersaCoT-10K and two public multimodal reasoning benchmarks, ScienceQA (Lu et al., 2022) and M^{3}CoT (Chen et al., 2024). Following (Yu et al., 2026), we adopt the corresponding RAD versions, which provide diverse reasoning states.

2) Implementation Details. We train all Qwen models for 5 epochs with a learning rate of $1e^{-5}$. Iterative self-refinement uses 9 rounds by default and reports peak accuracy. For settings without iterative refinement, the model performs one additional answer generation or correction round after initial CoT generation. All experiments are conducted on 8 NVIDIA H20 GPUs.

6.2 Comparison with SOTA Methods

1) Evaluation on the VersaCoT-10K dataset. As shown in Table 1, VersaCoT-10K exhibits clear performance gaps across MLLMs, with overall zero-shot accuracy ranging from 67.45% for Qwen2.5-

Table 2: Main results (%) on the ScienceQA dataset. NAT = natural science, SOC= social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. Red denotes the best result, and blue denotes the second best result.

Model	Learning	Size	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	**Avg**
Random	-	-	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67	39.83
Human	-	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
ViLT	*Fine-tune*	113M	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90	61.14
Patch-TRM	*Fine-tune*	90M	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50	61.42
VisualBERT	*Fine-tune*	111M	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92	61.87
LLaMA-Adapter	*Fine-tune*	6B	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
LLaVA	*Fine-tune*	13B	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LaVIN	*Fine-tune*	7B	89.25	94.94	85.24	88.51	87.46	88.08	90.16	88.07	89.41
Qwen2.5-VL-7B	*Zero-shot*	7B	83.79	88.30	82.45	83.63	84.38	81.81	86.67	80.29	84.39
Qwen3-VL-8B	*Zero-shot*	8B	90.54	94.49	87.64	92.18	91.67	86.20	90.64	90.57	90.62
Qwen3.5-9B	*Zero-shot*	9B	91.34	94.38	88.55	92.72	92.81	86.69	92.47	89.06	91.25
Qwen3.8-Max	*Zero-shot*	-	98.31	99.55	96.55	98.68	98.81	96.38	98.31	97.76	98.11
Grok-4.6	*Zero-shot*	-	95.60	95.39	94.64	96.63	94.89	93.80	96.15	93.80	95.31
GPT-5.6-Sol	*Zero-shot*	-	98.18	95.95	98.82	97.65	96.58	98.75	98.38	96.97	97.88
Gemini-3.5-Flash	*Zero-shot*	-	98.52	98.43	97.64	98.19	98.02	97.98	98.57	97.82	98.30
Claude Sonnet 5	*Zero-shot*	-	95.60	95.39	95.45	94.13	93.90	95.89	96.26	94.20	95.52
Claude Opus 5	*Zero-shot*	-	98.45	96.85	98.64	97.41	97.22	98.68	98.60	97.36	98.16
Seed-2.1-Pro	*Zero-shot*	-	97.11	97.98	96.27	97.41	96.68	96.24	97.10	97.03	97.08
Kimi-K3	*Zero-shot*	2.8T	97.96	98.31	99.09	98.00	97.57	98.82	98.46	98.09	98.33
Multimodal-CoT$_{base}$	*Fine-tune*	223M	84.06	92.35	82.18	82.75	82.75	84.74	85.79	84.44	85.31
MC-CoT$_{base}$	*Fine-tune*	223M	91.87	84.59	93.00	92.28	88.30	92.75	90.64	90.64	90.64
Multimodal-T-SciQ$_{base}$	*Fine-tune*	223M	91.52	91.45	92.45	91.94	90.33	92.26	92.11	91.10	91.75
MIND$_{base}$	*Fine-tune*	223M	93.07	96.74	87.09	92.42	92.76	89.27	92.91	91.17	92.29
Qwen2.5-VL-7B	*Fine-tune*	7B	97.02	97.86	92.82	96.77	96.33	94.15	96.92	94.66	96.11
Qwen2.5-VL-7B + Multimodal-CoT	*Fine-tune*	7B	97.47	97.19	90.82	97.46	96.58	92.40	96.04	95.06	95.68
Qwen2.5-VL-7B + MIND	*Fine-tune*	7B	97.56	98.20	93.36	97.56	97.03	94.49	97.06	95.78	96.60
Qwen2.5-VL-7B + ReState (Ours)	*Fine-tune*	7B	**97.91**	**99.44**	**94.82**	**98.04**	**98.17**	**95.54**	**97.91**	**96.57**	**97.43**
Qwen3-VL-8B	*Fine-tune*	8B	98.13	97.19	94.09	98.00	97.12	95.12	96.95	96.77	96.89
Qwen3-VL-8B + Multimodal-CoT	*Fine-tune*	8B	97.42	97.08	94.27	97.26	96.28	95.33	96.92	95.85	96.53
Qwen3-VL-8B + MIND	*Fine-tune*	8B	97.96	98.31	94.64	97.90	97.37	95.61	97.32	96.90	97.17
Qwen3-VL-8B + ReState (Ours)	*Fine-tune*	8B	**99.07**	**99.10**	**95.55**	**99.17**	**98.81**	**96.38**	**98.27**	**97.96**	**98.16**
Qwen3.5-9B	*Fine-tune*	9B	95.69	96.40	94.73	95.31	94.15	95.68	95.89	95.06	95.59
Qwen3.5-9B + Multimodal-CoT	*Fine-tune*	9B	96.40	96.63	92.73	96.29	95.49	94.01	95.78	94.99	95.50
Qwen3.5-9B + MIND	*Fine-tune*	9B	95.78	97.64	97.18	95.45	94.60	**97.63**	97.10	95.52	96.53
Qwen3.5-9B + ReState (Ours)	*Fine-tune*	9B	**98.71**	**98.76**	**97.36**	**98.78**	**98.46**	**97.63**	**98.60**	**97.96**	**98.37**

VL-7B (Bai et al., 2025b) to 86.85% for Grok-4.6 (xAI, 2026c). Performance also varies substantially across reasoning categories, and even frontier models remain unsaturated. These results establish VersaCoT-10K as a challenging and discriminative benchmark for fine-grained multimodal reasoning.

More importantly, under identical backbone and training settings, ReState achieves the best average performance on all three backbones, reaching 77.05%, 79.25%, and 81.90% with Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), and Qwen3.5-9B (Qwen Team, 2026a), respectively. Compared with direct fine-tuning, Multimodal-CoT (Zhang et al., 2024), and MIND (Yu et al., 2026), ReState improves accuracy by 1.90%-4.10%, 2.75%-4.60%, and 0.90%-1.70%, respectively. In addition, compared with the latest MIND, ReState also reduces training time by 31.2% (from 9.86 h to 6.78 h) and requires only one model to be deployed instead of two. Moreover, combining the results in Table 1 and Table 4, even without iterative self-refinement, i.e., under an inference cost comparable to MIND, ReState still improves accuracy by 0.15%-0.90% across the three backbones.

Notably, ReState with Qwen3.5-9B reaches 81.90%, outperforming substantially larger models such as Qwen3.6-27B (Qwen Team, 2026b), GLM-4.6V (106B) (Team et al., 2026b), MiniMax-M3 (428B) (Lai et al., 2026), and Kimi-K2.6 (1T) (Moonshot AI, 2026). Models that further surpass ReState are generally much larger. For example, the 2.8T-parameter Kimi-K3 (Kimi Team, 2026) reaches 85.65%. Although fine-tuning and zero-shot evaluation are not directly comparable, these results show that ReState effectively enhances the reasoning capability of medium-scale MLLMs. Its consistent gains across backbones and reasoning categories further support the effectiveness of state-centric reasoning. More comparisons and analyses are provided in *Appendix* E.1 and Table 8.

2) Evaluation on the ScienceQA dataset. As shown in Table 2, ReState achieves strong overall performance on ScienceQA. In particular, ReState with Qwen3.5-9B attains the best result of 98.37%, outperforming strong closed-source zero-shot models such as Qwen3.8-Max (98.11%) (Qwen Team, 2026d), Claude Opus 5 (98.16%) (Anthropic, 2026b), and Kimi-K3 (98.33%) (Kimi Team, 2026),

Table 3: Main results (%) on the M^3CoT dataset. The "Random" and "Human" results are averaged over three trials. Red denotes the best result, and blue denotes the second best result.

Model	Size	Science			Commonsense			Mathematics			**Total**
		Lang	Natural	Social	Physical	Social	Temporal	Algebra	Geometry	Theory	
Random	-	32.70	30.62	26.71	32.97	22.22	20.33	35.71	27.50	23.81	28.56
Human	-	97.83	92.62	94.31	96.28	92.41	88.71	87.23	88.75	85.71	91.61
Tool-Usage Methods											
HuggingGPT	175B	17.57	20.93	10.33	8.70	14.75	9.76	11.35	22.50	9.52	14.60
VisualChatGPT	>175B	30.09	36.28	7.78	43.48	29.92	33.33	21.99	21.25	28.57	25.92
Chameleon	-	43.87	26.05	25.44	39.13	37.30	48.78	17.73	26.25	23.81	34.29
Zero-shot Methods											
Qwen2.5-VL-7B	7B	83.89	59.77	47.45	81.11	76.03	79.67	41.43	30.00	28.57	59.79
Qwen3-VL-8B	8B	86.73	65.77	45.86	82.22	77.27	77.24	34.29	32.50	38.10	61.43
Qwen3.5-9B	9B	89.57	66.28	47.45	87.78	80.99	78.86	30.71	33.75	38.10	62.81
Qwen3.8-Max	-	94.31	83.27	67.68	90.00	82.23	94.31	61.43	60.00	57.14	78.43
Grok-4.6	-	93.84	85.19	73.89	91.11	75.62	94.31	70.71	72.50	53.38	81.02
GPT-5.6-Sol	-	98.10	78.80	66.08	88.89	79.75	91.06	73.57	48.75	61.90	76.75
Gemini-3.5-Flash	-	99.05	85.95	72.45	93.33	85.12	95.12	78.57	63.75	85.71	82.96
Claude Sonnet 5	-	95.73	76.37	57.01	87.78	79.75	90.24	59.29	37.50	52.38	71.83
Claude Opus 5	-	97.63	83.65	65.45	90.00	80.17	95.12	88.57	66.25	61.90	79.98
Seed-2.1-Pro	-	96.68	72.67	65.13	88.89	78.93	89.43	68.57	56.25	28.57	73.77
Kimi-K3	2.8T	97.16	86.72	70.38	91.11	78.51	95.93	77.86	60.00	57.14	81.32
Fine-tuning Methods											
LLaVA-V1.5	13B	68.72	72.41	40.86	83.52	64.61	69.11	35.71	45.00	38.10	59.50
Multimodal-CoT$_{base}$	223M	41.71	46.49	39.90	59.34	60.91	27.64	48.57	35.00	28.57	44.85
MC-CoT$_{base}$	223M	53.55	63.98	43.56	61.54	69.55	29.27	42.86	33.75	28.57	53.51
MIND$_{base}$	223M	72.04	63.09	43.31	82.22	65.29	44.72	49.29	61.25	33.33	57.38
Qwen2.5-VL-7B	7B	92.42	91.83	60.35	84.44	**80.17**	86.18	72.14	70.00	71.43	79.42
Qwen2.5-VL-7B + Multimodal-CoT	7B	87.20	89.27	61.62	83.33	74.38	86.18	56.43	61.25	47.62	76.32
Qwen2.5-VL-7B + MIND	7B	92.42	**97.32**	**76.27**	83.33	75.62	**88.62**	66.43	71.25	61.90	84.81
Qwen2.5-VL-7B + ReState (Ours)	7B	**94.79**	**97.32**	71.82	**85.56**	78.10	86.18	**75.71**	**78.75**	**76.19**	**84.99**
Qwen3-VL-8B	8B	93.84	90.55	68.31	**86.67**	80.99	89.43	74.29	67.50	66.67	81.62
Qwen3-VL-8B + Multimodal-CoT	8B	90.05	92.72	65.92	83.33	67.36	90.24	62.14	62.50	66.67	78.95
Qwen3-VL-8B + MIND	8B	94.79	97.96	**79.62**	85.56	80.17	**95.12**	68.57	**77.50**	66.67	87.45
Qwen3-VL-8B + ReState (Ours)	8B	**95.73**	**98.47**	78.03	83.33	**83.88**	92.68	**76.43**	76.25	**76.19**	**87.96**
Qwen3.5-9B	9B	93.36	96.42	74.52	85.56	80.58	84.55	70.00	71.25	61.90	84.73
Qwen3.5-9B + Multimodal-CoT	9B	91.00	89.66	59.87	85.56	67.77	82.93	60.00	70.00	33.33	75.93
Qwen3.5-9B + MIND	9B	95.26	**98.08**	**81.69**	85.56	81.40	85.37	72.86	**83.75**	57.14	88.09
Qwen3.5-9B + ReState (Ours)	9B	**95.73**	97.83	80.89	**88.89**	**83.47**	**86.99**	**80.00**	**83.75**	**90.48**	**89.00**

as well as existing learning-based MCoT methods (Zhang et al., 2024; Tan et al., 2024a; Wang et al., 2024a; Yu et al., 2026). Under identical backbone and training settings, ReState consistently outperforms direct SFT, Multimodal-CoT (Zhang et al., 2024), and MIND (Yu et al., 2026) by 1.27%-2.78%, 1.63%-2.87%, and 0.83%-1.84%, respectively. Moreover, ReState achieves the best results across all eight sub-categories. Overall, these results show the consistent effectiveness of ReState across MLLMs. More comparisons and analyses are provided in *Appendix* E.2 and Table 9.

3) Evaluation on the M^3CoT dataset. As shown in Table 3, ReState achieves 84.99%, 87.96%, and 89.00% with Qwen2.5-VL-7B, Qwen3-VL-8B, and Qwen3.5-9B, respectively. In particular, ReState with Qwen3.5-9B attains the best result of 89.00%, outperforming strong zero-shot models such as Gemini-3.5-Flash (82.96%) (Google DeepMind, 2026), Kimi-K3 (81.32%) (Kimi Team, 2026), and Grok-4.6 (81.02%) (xAI, 2026c), while trailing human performance (91.61%) by only 2.61%. Under identical backbone settings, ReState improves over direct SFT, Multimodal-CoT (Zhang et al., 2024), and MIND (Yu et al., 2026) by 4.27%–6.34%, 8.67%–13.07%, and 0.18%–0.91%, respectively, demonstrating consistent gains across different base models. It also achieves the best results on 7 of 9 sub-tasks, with notable gains on Theory (from 61.90% to 90.48%) and Algebra (from 70.00% to 80.00%). These results show that ReState provides consistent benefits across diverse knowledge structures and reasoning patterns, with particularly clear advantages on more challenging multimodal reasoning tasks. More comparisons and analyses are provided in *Appendix* E.3 and Table 10.

6.3 Break-Down Ablation

We conduct progressive ablations on VersaCoT-10K across three backbones to evaluate the contributions of URSM and CRSL. From Table 4, URSM alone improves the baseline by 0.50% (from 75.15% to 75.65%)-2.55% (from 75.65% to 78.20%), demonstrating the benefit of unifying reasoning construction, preservation, and correction as state transitions within the same MLLM. Since both training-time self-history state (SHS) and inference-time iterative self-refinement (ISR) in CRSL build upon URSM, we further ablate SHS and ISR on top of URSM. Adding training-time SHS yields an additional 0.30% (from 78.20% to 78.50%) - 2.15% (from 78.95% to 81.10%) gain, indicating

Table 4: The Break-Down ablation of ReState on the VersaCoT-10K dataset. SHS and ISR denote the training-time self-history state mechanism and inference-time iterative self-refinement in CRSL.

Scheme	URSM	CRSL		Qwen2.5-VL-7B	Qwen3-VL-8B	Qwen3.5-9B
		SHS	ISR			
Baseline	✗	✗	✗	75.15 (+1.90)	75.65 (+3.60)	77.80 (+4.10)
Baseline w/ URSM	✓	✗	✗	75.65 (+1.40)	78.20 (+1.05)	78.95 (+2.95)
Baseline w/ URSM, CRSL(SHS)	✓	✓	✗	76.20 (+0.85)	78.50 (+0.75)	81.10 (+0.80)
Baseline w/ URSM, CRSL(ISR)	✓	✗	✓	76.20 (+0.85)	78.65 (+0.60)	79.45 (+2.45)
ReState (Final)	✓	✓	✓	**77.05**	**79.25**	**81.90**

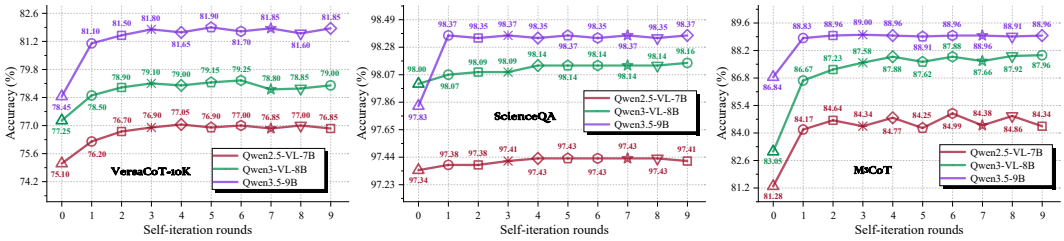


Figure 4: Performance evolution over iterative self-refinement rounds across different MLLMs and datasets. Round 0 denotes the initial reasoning, with most gains achieved in the first few rounds.

that self-generated historical states help the model better adapt to model-induced states. Adding inference-time ISR further improves performance by 0.45% (78.20% to 78.65%) - 0.55% (from 75.65% to 76.20%), validating the effectiveness of recursive state feedback for refining existing reasoning. Combining URSM with both SHS and ISR achieves the best performance across all three backbones, improving the respective baselines by 1.90%, 3.60%, and 4.10%. Overall, URSM provides unified state modeling, while SHS reduces the training-inference reasoning-state gap and ISR enables recursive refinement. Their complementary gains support the effectiveness of the proposed state-centric reasoning framework. More experiments and analyses are provided in *Appendix* D.

6.4 Discussion of "Iterative Self-Refinement"

To further analyze the iterative self-refinement capability of ReState at inference, we evaluate its performance under different numbers of refinement rounds in Fig. 4. Across all three backbones and datasets, ReState consistently benefits from iterative self-refinement. After only one refinement round, performance improves by 1.10%-2.65% on VersaCoT-10K, 0.04%-0.54% on ScienceQA, and 1.99%-3.62% on M^{3}CoT. Compared with round 0, the best gains reach 1.95%-3.45%, 0.09%-0.54%, and 2.16%-4.91%, respectively. Performance generally improves rapidly in early rounds and then stabilizes with minor fluctuations, with larger gains on the more challenging VersaCoT-10K and M^{3}CoT. More importantly, ReState reuses the same MLLM throughout the refinement process and requires no additional reflection module, critic, or independent correction model. Instead, refinement is achieved through a recursive loop of "state feedback → re-evaluation → state update". These results validate its ability to iteratively refine explicit reasoning states during inference, further supporting the effectiveness of state-centric reasoning. More discussions are provided in *Appendix* G.

7 Conclusion

This work revisits the stage-centric paradigm of learning-based MCoT and identifies two key limitations: fragmented modeling of reasoning behaviors and the training-inference reasoning-state gap. Therefore, we propose ReState, a unified iterative reasoning framework that reformulates multimodal reasoning as dynamic state evolution. URSM unifies reasoning construction, preservation, and correction within a shared state-update mechanism, while CRSL incorporates historical-state learning and recursive feedback to reduce the distribution gap and enable iterative self-refinement without additional reflection or correction modules. We also construct VersaCoT-10K, a high-quality CoT dataset spanning eight visual reasoning domains. Extensive experiments demonstrate the consistent effectiveness and cross-model generalizability of ReState. We hope this work can draw attention to the research on state-centric multimodal reasoning and advance MLLMs toward more general, reliable, and adaptive reasoning systems.

AI USE STATEMENT

We used generative AI tools for manuscript polishing and several components of the data construction and evaluation pipelines. Specifically, generative AI was used to assist with language polishing. For VersaCoT-10K, Gemini-3-Flash was used to generate candidate CoT annotations, Gemini-3.5-Flash was used as a multimodal quality judge to filter generated rationales, and Qwen2.5-VL-72B was used to generate image captions. Generative models were used as assisting tools rather than as substitutes for author verification. We reviewed the AI-assisted writing, data-processing procedures, generated annotations, and reported results, and take full responsibility for the final content of this work, including all text, data, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We support reproducibility by detailing the ReState framework in Section 4 of the main text. Appendix A provides the complete construction pipeline of VersaCoT-10K, including data collection and standardization, CoT annotation and quality control, and reasoning-state augmentation. Appendix C details the experimental setup. ***Notably, the supplementary material provides our anonymized training and inference code.*** Together, these materials provide the methodological, data-processing, and implementation details needed to reproduce the reported results.

REFERENCES

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

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6077–6086, 2018.

Anthropic. Claude haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5, October 2025.

Anthropic. Claude opus 4.8. https://www.anthropic.com/news/claude-opus-4-8, May 2026a.

Anthropic. Claude opus 5. https://www.anthropic.com/news/claude-opus-5, July 2026b.

Anthropic. Claude sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6, February 2026c.

Anthropic. Claude sonnet 5. https://www.anthropic.com/news/claude-sonnet-5, June 2026d.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, et al. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, et al. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19553–19564. IEEE, 2025.

ByteDance Seed. Introduction to techniques used in seed1.6. https://seed.bytedance.com/en/blog/introduction-to-techniques-used-in-seed1-6, June 2025. Published June 25, 2025.

ByteDance Seed. Seed2.1: A next-generation agent for real-world productivity. https://seed.bytedance.com/en/seed2_1, June 2026.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 513–523, 2021.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8199–8221, 2024.

Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. *Advances in neural information processing systems*, 38:69110–69139, 2026.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Hao Zhang, and Chuang Gan. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning. *arXiv preprint arXiv:2301.05226*, 2023.

Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23678–23686, 2025.

Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi Wang, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, et al. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *Advances in Neural Information Processing Systems*, 38: 96084–96112, 2026.

Charles Corbière, Simon Roburin, Syrielle Montariol, Antoine Bosselut, and Alexandre Alahi. Drivingvqa: A dataset for interleaved visual chain-of-thought in real-world driving scenarios. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3309–3333, 2026.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 49250–49267, 2023.

Qihua Dong, Yitian Zhang, Huimin Zeng, Yizhou Wang, Jianglin Lu, Kuo Yang, and Yun Fu. Revealing the seen, imagining the beyond: A survey of image-grounded chain-of-thought reasoning in multimodal llms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45055–45070, 2026.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9062–9072. IEEE, 2025.

Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6639–6648, 2019.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. In *International Conference on Machine Learning*, pp. 11398–11442. PMLR, 2023.

Google DeepMind. Gemini 3 flash model card. https://deepmind.google/models/model-cards/gemini-3-flash/, December 2025. Published December 17, 2025.

Google DeepMind. Gemini 3.5 flash model card. https://deepmind.google/models/model-cards/gemini-3-5-flash/, May 2026. Published May 19, 2026.

Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. Multi-modal latent space learning for chain-of-thought reasoning in language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 18180–18187, 2024.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Zhe Xu, Yao Hu, Shaohui Lin, et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 63794–63812, 2026.

Zhuoran Jin, Kejian Zhu, Hongbang Yuan, Yupu Hao, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Look light, think heavy: What multimodal chain-of-thought reasoning can and cannot do. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8567–8608, 2026.

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

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*, 2020.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 5583–5594. PMLR, 2021.

Kimi Team. Kimi k3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Xunhao Lai, Weiqi Xu, Yufeng Yang, Qiaorui Chen, Yang Xu, Lunbin Zeng, Xiaolong Li, Haohai Sun, Haichao Zhu, Vito Zhang, Jinkai Hu, Jiayao Li, Rui Gao, Zekun Li, Songquan Zhu, Jingkai Zhou, and Pengyu Zhao. Minimax sparse attention, 2026. URL https://arxiv.org/abs/2606.13392.

Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. Multimodal reasoning with multimodal knowledge graph. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10767–10782, 2024.

Chuhan Li, Ziyao Shangguan, Yilun Zhao, Deyuan Li, Yixin Liu, and Arman Cohan. M3sciqa: A multi-modal multi-document scientific qa benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15419–15446, 2024.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 5265–5275, 2020.

Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, Xuan-Jing Huang, and Zhongyu Wei. Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. 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)*, pp. 3769–3798, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, 2024.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36:43447–43478, 2023.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:29615–29627, 2023.

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Argus: Vision-centric reasoning with grounded chain-of-thought. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14268–14280. IEEE, 2025.

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

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, et al. Chartqapro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19123–19151, 2025.

Fanxu Meng, Haotong Yang, Yiding Wang, and Muhan Zhang. Chain of images for intuitively reasoning. *arXiv preprint arXiv:2311.09241*, 2023.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14420–14431. IEEE, 2024.

Moonshot AI. Kimi k2.6. https://www.kimi.com/ai-models/kimi-k2-6, 2026.

OpenAI. Gpt-4v(ision) system card. https://openai.com/index/gpt-4v-system-card/, September 2023. Published September 25, 2023.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

OpenAI. Gpt-5.4 thinking system card. https://openai.com/index/gpt-5-4-thinking-system-card/, March 2026a. Published March 5, 2026.

OpenAI. Gpt-5.5 system card. https://openai.com/index/gpt-5-5-system-card/, April 2026b. Published April 23, 2026.

OpenAI. Gpt-5.6 system card. https://deploymentsafety.openai.com/gpt-5-6, July 2026c. Published July 9, 2026.

OpenAI. Openai gpt-5 system card, 2026d. URL https://arxiv.org/abs/2601.03267.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026c. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026d. URL https://qwen.ai/blog?id=qwen3.8.

Qwen Team. Qwen3.6-Plus: Towards real world agents, April 2026e. URL https://qwen.ai/blog?id=qwen3.6.

Qwen Team. Qwen3.7: The agent frontier, May 2026f. URL https://qwen.ai/blog?id=qwen3.7.

Bytedance Seed. Seed1.8 model card: Towards generalized real-world agency, 2026a. URL https://arxiv.org/abs/2603.20633.

Bytedance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity, 2026b. URL https://arxiv.org/abs/2607.00248.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:38154–38180, 2023.

Cheng Tan, Jingxuan Wei, Zhangyang Gao, Linzhuang Sun, Siyuan Li, Ruifeng Guo, Bihui Yu, and Stan Z Li. Boosting the power of small multimodal reasoning models to match larger models with self-consistency training. In *European Conference on Computer Vision*, pp. 305–322. Springer, 2024a.

Cheng Tan, Jingxuan Wei, Linzhuang Sun, Zhangyang Gao, Siyuan Li, Bihui Yu, Ruifeng Guo, and Stan Z Li. Retrieval meets reasoning: Even high-school textbook knowledge benefits multimodal reasoning. *arXiv preprint arXiv:2405.20834*, 2024b.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. *Advances in neural information processing systems*, 38:5772–5822, 2026.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

V Team, Wenyi Hong, Xiaotao Gu, Ziyang Pan, Zhen Yang, Yuting Wang, Yue Wang, Yuanchang Yue, Yu Wang, Yanling Wang, Yan Wang, Xijun Liu, et al. Glm-5v-turbo: Toward a native foundation model for multimodal agents, 2026a. URL https://arxiv.org/abs/2604.26752.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, et al. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2026b. URL https://arxiv.org/abs/2507.01006.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37: 87310–87356, 2024.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19162–19170, 2024a.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:121475–121499, 2024b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of llms: visualization-of-thought elicits spatial reasoning in large language models. *Advances in Neural Information Processing Systems*, 37:90277–90317, 2024a.

Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Philip Torr, and Jian Wu. Dettoolchain: A new prompting paradigm to unleash detection ability of mllm. In *European Conference on Computer Vision*, pp. 164–182. Springer, 2024b.

xAI. Grok 4.3. `https://docs.x.ai/developers/models/grok-4.3`, 2026a.

xAI. Introducing grok 4.5. `https://x.ai/news/grok-4-5`, July 2026b.

xAI. Introducing grok 4.6. `https://x.ai/news/grok-4-6`, August 2026c.

Zhengzhuo Xu, Bowen Qu, Yiyan Qi, Sinan Du, Chengjin Xu, Chun Yuan, and Jian Guo. Chartmoe: Mixture of diversely aligned expert connector for chart understanding. In *International Conference on Learning Representations*, volume 2025, pp. 78550–78572, 2025.

Cheng Yang, Chufan Shi, Yaxin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran Xu, Xinyu Zhu, Siheng Li, Yuxiang Zhang, et al. Chartmimic: Evaluating lmm's cross-modal reasoning capability via chart-to-code generation. In *International Conference on Learning Representations*, volume 2025, pp. 26590–26646, 2025a.

Shuo Yang, Caren Han, Siwen Luo, and Eduard Hovy. Magic-vqa: Multimodal and grounded inference with commonsense knowledge for visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16967–16986, 2025b.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11289–11303, 2023.

Chuang Yu, Jinmiao Zhao, Mingxuan Zhao, Yunpeng Liu, Xiujun Shu, Yuanhao Feng, Bo Wang, and Xiangyu Yue. Mind: Multi-rationale integrated discriminative reasoning framework for multi-modal large models. In *Forty-third International Conference on Machine Learning*, 2026.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6281–6290, 2019.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 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*, pp. 9556–9567, 2024.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023a.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Shanghang Zhang, Gao Peng, et al. Mavis: Mathematical visual instruction tuning with an automatic data engine. In *International Conference on Learning Representations*, volume 2025, pp. 87955–87989, 2025.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023b.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024, 2024.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4995–5004, 2016.

A Detailed Construction for VersaCoT-10K

As shown in Fig. 3, VersaCoT-10K is constructed through three consecutive stages: *Data Collection and Standardization*, *CoT Annotation and Quality Control*, and *Reasoning-State Augmentation*. We aim to integrate reasoning-intensive samples spanning diverse visual content, domain knowledge, and reasoning patterns into a unified single-image multiple-choice format, providing high-quality supervision for multi-domain reasoning and reasoning-state refinement. For each of the eight heterogeneous visual reasoning domains, we retain 1,250 samples and split them into training, validation, and test sets at a 7:1:2 ratio. VersaCoT-10K therefore contains 10,000 samples in total, including 7,000 for training, 1,000 for validation, and 2,000 for testing. The following subsections detail each stage of the construction pipeline.

A.1 Data Collection and Standardization

To ensure broad coverage of multimodal reasoning, we collect reasoning-intensive samples with explicit visual inputs and valid answer annotations from multiple public benchmarks. According to their primary visual content and reasoning characteristics, these samples are organized into eight heterogeneous visual reasoning domains: medical reasoning, driving and safety, visual commonsense, scientific diagram, spatial perception, geometric and mathematical, chart computation, and multi-disciplinary knowledge. The corresponding source datasets are PMC-VQA (Zhang et al., 2023b), DrivingVQA (Corbière et al., 2026), Visual7W (Zhu et al., 2016), AI2D (Kembhavi et al., 2016), CV-Bench (Tong et al., 2024), GeoQA (Chen et al., 2021), ChartQA (Masry et al., 2022), and MMMU (Yue et al., 2024), respectively. Together, these datasets provide diverse visual evidence, domain knowledge, and reasoning patterns. Because the source datasets differ substantially in task formulation, answer representation, and annotation structure, we further standardize the collected samples into a unified single-image multiple-choice format. As illustrated in Fig. 3, this process consists of three steps: *Option Construction*, *Answer Normalization*, and *Input Format Organization*.

Option Construction. For source datasets that already adopt a multiple-choice format, we preserve the original candidate options whenever possible while standardizing their textual representation. For datasets originally formulated as open-ended question answering, we construct discriminative distractors from the available annotations and associated structured information, thereby converting them into a unified multiple-choice format. For ChartQA in particular, originally free-form Yes/No questions are directly converted into binary-choice questions. For questions whose answers correspond to textual entries, years, or numerical values, distractors are preferentially sampled from the structured table associated with the chart. For numerical questions, we first select table values with the same numerical format as the correct answer. When an insufficient number of valid candidates is available, limited deterministic numerical perturbations are applied to construct additional distractors. Samples are discarded when reliable and discriminative incorrect options cannot be constructed. The resulting candidate set for sample i is represented as

$$\mathcal{O}_i = \{o_{i,1}, \ldots, o_{i,n_i}\}, \tag{6}$$

where n_i denotes the number of candidate options. We preserve the original variation in candidate-set size across source datasets rather than forcing all samples into a fixed answer space.

Answer Normalization. The source datasets encode correct answers in heterogeneous forms, including letter labels, numerical indices, and text strings. To establish a unified training and evaluation interface, we normalize the correct answer of each sample to its zero-based index in the standardized candidate set:

$$\mathcal{A}_i \in \{0, \ldots, n_i - 1\}. \tag{7}$$

For answers represented by letter or numerical labels, we first recover their positions according to the original option ordering and then map them to the standardized candidate set. Textual answers are matched against the standardized options before being converted into the corresponding indices. We additionally verify that each answer uniquely corresponds to exactly one candidate option, ensuring that every retained sample has a unique and valid ground-truth answer.

Input Format Organization. After option construction and answer normalization, we further standardize the multimodal input structure across datasets. Each sample i is represented as

$$\mathcal{D}_i = \{\mathcal{I}_i,\ \mathcal{Q}_i,\ \mathcal{O}_i,\ \mathcal{C}_i,\ \mathcal{A}_i\}, \tag{8}$$

You are an expert annotator for a single-image multiple-choice VQA dataset.
Given one image, one question, and several answer options, write a concise, image-grounded reasoning annotation that supports the correct option.

Dataset information:
- category: {*category_id*} {*category*}

Output format:
Return exactly one <think>...</think> block followed by one <answer>...</answer> block.
Do not output markdown, JSON, headings, bullet lists, or extra text outside the tags.
Keep the reasoning inside the <think>...</think> block within 1000 tokens.
The <answer> block must contain only the final option in this format: <answer>(A) answer text</answer>.

Reasoning rules:
1. Ground the <think> block in the image, question, and choices.
2. Reason forward from visual evidence to the conclusion; do not write an answer-driven justification.
3. Do not mention that the answer, annotation, label, target, or constraint was provided.
4. Use sample-specific visual details, not generic explanations.
5. Compare plausible choices when helpful, and explain why distractors are less supported.
6. Do not invent visual details or unsupported domain claims.
7. For counting, spatial reasoning, diagrams, charts, geometry, or math, include the key intermediate logic or calculation.
8. Keep the rationale logically complete, while avoiding unnecessary verbosity.

Category guidance:
- Medical: rely only on visible medical evidence and avoid unsupported diagnostic or treatment claims.
- Driving/safety: focus on road layout, traffic objects, signs, obstacles, and spatial safety cues.
- Natural scenes: combine visible evidence with minimal necessary commonsense.
- Scientific diagrams: use labels, arrows, components, processes, and symbolic relations.
- Spatial/perceptual tasks: focus on objects, counts, positions, orientation, depth, and occlusion.
- Geometry/math: state the relevant relation, shape, measurement, formula, or calculation.
- Charts/tables: use axes, legends, values, trends, comparisons, and necessary arithmetic.
- Expert-domain questions: use domain knowledge only when grounded in the image.

Question:
{*question*}

Choices:
{*choices*}

Final-answer consistency:
The final <answer> must be exactly:
<answer>{*gold_answer*}</answer>

Use this only to keep the annotation consistent with the dataset label. The <think> must remain a natural, image-grounded, forward reasoning process. Do not mention that the answer was supplied or constrained.

Now produce only:\n<think>...</think>\n<answer>...</answer>

Figure 5: Prompt Design for CoT Annotation during the construction of the VersaCoT-10K dataset. Italicized text represents assignment variables.

where $\mathcal{I}_i$, $\mathcal{Q}_i$, $\mathcal{O}_i$, $\mathcal{C}_i$, and $\mathcal{A}_i$ denote the input image, question, candidate-option set, image caption, and correct answer index, respectively. We standardize image storage, question fields, and sample identifiers, and retain only samples compatible with single-image reasoning. Additional task-specific filtering is applied when necessary. For Visual7W, we retain only samples from the *telling* subset that are compatible with text-based multiple-choice reasoning. For MMMU we retain only samples satisfying the single-image multiple-choice requirement. We also preserve metadata such as the source dataset and original split to support subsequent partitioning and statistical analysis. In addition, for source datasets with predefined training and test partitions, we maintain source-level split consistency during reorganization to avoid mixing samples from their original training and test sets. Through

You are a strict multimodal quality judge for a multi-domain single-image multiple-choice VQA dataset. Given one image, one question, answer choices, the annotated answer, and a generated rationale, judge whether the rationale is correct and image-grounded.

Dataset information:
- category: {*category_id*} {*category*}

Accept only if:
1. The rationale is grounded in the image, question, and choices.
2. The rationale reasons forward from visual evidence or necessary calculation to the annotated answer.
3. The rationale supports the annotated answer with sample-specific evidence, not generic wording or answer-only restatement.
4. The rationale does not hallucinate visual details, unsupported domain facts, or leak that the gold answer/label/constraint was provided.
5. For counting, spatial, diagram, chart, table, geometry, or math tasks, the key comparison or calculation is correct.

Category notes:
- Medical: reject unsupported diagnosis, treatment, or overclaiming.
- Driving/safety: check road layout, signs, vehicles, pedestrians, obstacles, and spatial safety cues.
- Natural scenes: allow commonsense only when supported by the image.
- Diagrams/charts/math: check labels, arrows, components, axes, values, trends, formulas, and calculations.
- Spatial/perceptual: check objects, counts, positions, orientation, depth, and occlusion.
- Expert-domain: allow domain knowledge only when tied to visible evidence.

Reject if the rationale is visually unsupported, incomplete, contradictory, generic, hallucinated, answer-leaking, or too weak to justify the answer. If uncertain, reject.

Scores use 0-10:
10 = excellent, 8-9 = good, 6-7 = partial, 4-5 = weak, 1-3 = poor, 0 = invalid.

Return strict JSON only:
{{"accept": true or false,
"reasoning_correct": true or false,
"reasoning_supports_answer": true or false,
"visual_grounding_score": 0-10,
"reasoning_correctness_score": 0-10,
"reasoning_completeness_score": 0-10,
"answer_support_score": 0-10,
"specificity_score": 0-10,
"failure_reasons": ["visual_hallucination" | "missing_key_evidence" | "contradiction" | "unsupported_jump" | "generic_template" | "answer_leakage" | "format_error" | "incomplete_reasoning" | "wrong_counting_logic" | "wrong_spatial_logic" | "wrong_chart_calculation" | "wrong_geometry_math_logic" | "unsupported_medical_claim" | "too_short" | "too_verbose" | "other"],}}

Question:
{*question*}

Choices:
{*choices*}

Annotated answer:
{*gold_answer*}

Generated rationale:
{*solution*}

Figure 6: Prompt Design for CoT filtering during the construction of the VersaCoT-10K dataset. Italicized text represents assignment variables. A quality score of 8 or higher is deemed acceptable.

these steps, heterogeneous source datasets are transformed into a unified single-image multiple-choice format for subsequent CoT annotation and reasoning-state construction.

A.2 CoT Annotation and Quality Control

After data standardization, we construct explicit reasoning annotations through *CoT Annotation and Quality Control*. Since most source visual question answering datasets do not provide high-quality explicit CoTs, they cannot directly support reasoning-process learning and subsequent reasoning-state construction. We therefore employ an automated MLLM-assisted annotation pipeline to generate visually grounded reasoning trajectories and apply strict quality control to retain reliable annotations. Before annotation, samples within each domain-specific data pool are randomly shuffled to reduce the dependence of final sample selection on the original dataset ordering.

CoT Annotation. Manually annotating reasoning trajectories for 10K samples spanning multiple visual reasoning domains would require substantial human effort. We therefore employ Gemini-3-Flash (Google DeepMind, 2025) as the CoT generator. Given the task information and a predefined prompt, the model produces a reasoning trajectory $\mathcal{R}_i$ together with a predicted answer $\hat{\mathcal{A}}_i$:

$$(\mathcal{R}_i, \hat{\mathcal{A}}_i) = \mathcal{M}_{gen}(\mathcal{D}_i, \mathcal{P}_{gen}), \tag{9}$$

where $\mathcal{P}_{gen}$ denotes the CoT generation prompt, provided in full in Fig. 5.

To construct label-consistent and visually grounded reasoning annotations, $\mathcal{P}_{gen}$ provides the source benchmark answer as the target conclusion and explicitly constrains how that conclusion should be justified. In particular, the generated rationale must establish a step-by-step reasoning path from concrete visual evidence and task-relevant information, without referring to or revealing the provided answer, gold label, dataset annotation, or any equivalent supervision signal. It must also avoid introducing visual evidence unsupported by the input image. For reasoning-intensive tasks such as counting, spatial reasoning, chart understanding, scientific diagrams, geometry, and mathematics, the rationale is further required to include the key comparisons, relational judgments, or intermediate calculations necessary to support the final conclusion.

To standardize the annotation format, the generator is required to produce outputs following

$$< think > \ reasoning\ process \ < /think > \ < answer > final\ option \ < /answer > \tag{10}$$

The generated answer $\hat{\mathcal{A}}_i$ is constrained to match the original annotation $\mathcal{A}_i^{GT}$ for label consistency. We then extract the `<think>` content as the candidate CoT annotation $\mathcal{R}_i$.

Quality Control. All generated CoTs undergo a two-stage quality-control procedure consisting of *rule-based filtering* and *judge-based filtering*. Rule-based filtering first verifies output structure, field validity, answer consistency, and text parsability. We restrict the reasoning length to 200-4,000 characters to remove overly short rationales with insufficient reasoning content as well as abnormally long outputs. We further detect answer-leakage expressions, such as "provided answer", "given answer", "gold answer", "target answer", and "dataset label". Samples whose rationales explicitly reveal dependence on externally supplied answer annotations are discarded.

Satisfying these structural constraints alone does not guarantee that a generated CoT is visually grounded or logically reliable. We therefore employ Gemini-3.5-Flash (Google DeepMind, 2026), which is separate from the CoT generator, as a multimodal quality judge. The judge temperature is set to 0 to reduce sampling variability. Given the original task information $\mathcal{D}_i$, the generated reasoning trajectory $\mathcal{R}_i$, and a predefined filtering prompt $\mathcal{P}_{filter}$, the judge evaluates the visual grounding, logical validity, and answer-supporting quality of each candidate CoT. Importantly, the judge is used for reasoning-annotation quality control rather than independent answer verification. Since the source benchmarks already provide validated target answers, the judge assesses whether the generated rationale provides a visually grounded and logically coherent justification for the annotated target. In this way, the source benchmark specifies the target answer, while the multimodal judge evaluates the quality of the reasoning supporting that target.

For each candidate, the judge produces three binary quality judgments and five fine-grained scores. The binary judgments indicate whether the rationale is *overall acceptable*, whether the *reasoning is correct*, and whether the *reasoning supports the answer*. The five scores evaluate *visual grounding*, *reasoning correctness*, *reasoning completeness*, *answer support*, and *specificity*, respectively, each on a scale from 0 to 10. For sample i, we denote the resulting score vector as

$$\mathbf{s}_i = [s_i^{vg}, s_i^{rc}, s_i^{comp}, s_i^{as}, s_i^{spec}]. \tag{11}$$

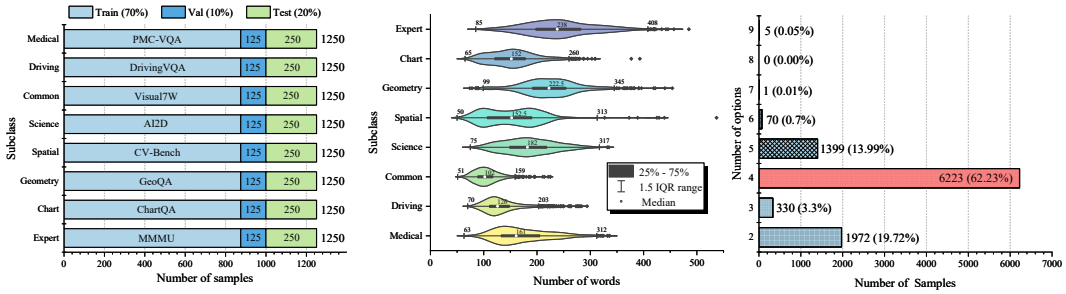


Figure 7: Statistics of VersaCoT-10K dataset. **Left:** data sources and splits across eight reasoning categories. **Middle:** CoT length distribution. **Right:** distribution of candidate-option counts.

Let c_i^{acc}, c_i^{cor}, and c_i^{sup} denote the three binary judgments. We retain a sample only when all three judgments are positive and every fine-grained score is at least 8:

$$\mathcal{D}_{keep} = \{\mathcal{D}_i \mid c_i^{acc} = c_i^{cor} = c_i^{sup} = 1,\ \min(\mathbf{s}_i) \geq 8\}\,. \tag{12}$$

The complete multimodal judging prompt is provided in Fig. 6. Through this combination of rule-based filtering and multimodal quality assessment, we retain quality-controlled CoT annotations that satisfy stringent criteria for visual grounding, logical coherence, answer support, and specificity, providing reliable supervision for subsequent reasoning-state augmentation.

A.3 Reasoning-State Augmentation

After CoT Annotation and Quality Control, each retained sample is associated with a high-quality correct reasoning trajectory. However, supervision from a single correct CoT alone cannot expose the model to the diverse reasoning-state conditions required for reasoning-state understanding and refinement. We therefore adopt the Rationale Augmentation and Discrimination (RAD) paradigm (Yu et al., 2026) to construct diverse positive states and hard negative states from each filtered CoT.

Specifically, given a correct CoT, we construct positive reasoning states $\mathcal{R}_i^+$ by introducing variations in wording, organization, and selected reasoning details while preserving the original reasoning semantics and conclusion. These diverse but semantically consistent states expose the model to different valid realizations of the same reasoning process, providing supervision for preserving and refining correct reasoning. In parallel, we construct hard negative reasoning states $\mathcal{R}_i^-$ by modifying a small number of critical semantic elements, visual relations, or inference steps. Such negative states remain superficially similar to the correct reasoning trajectory while containing key reasoning errors, thereby providing challenging supervision for identifying and correcting erroneous reasoning.

Together with the empty state ϕ, the pre-constructed reasoning-state set for sample i is defined as

$$\mathcal{S}_i^{RAD} = \left\{\phi,\ \mathcal{R}_i^+,\ \mathcal{R}_i^-\right\}, \tag{13}$$

where ϕ is represented by the special token `[EMPTY_SOLUTION]`. These three state types correspond to the three reasoning behaviors modeled by URSM: the empty state ϕ provides supervision for reasoning construction, positive states $\mathcal{R}_i^+$ for correct reasoning preservation and refinement, and negative states $\mathcal{R}_i^-$ for erroneous reasoning correction. During training, positive and negative states are sampled with equal probability when a non-empty reasoning state is selected.

It is important to distinguish these RAD-generated states from the historical reasoning states used by CRSL. The positive and negative states $\mathcal{R}_i^+$ and $\mathcal{R}_i^-$ are pre-constructed dataset-level annotations and remain fixed throughout training. In contrast, the historical states $\mathcal{R}_i^h$ are not part of the static VersaCoT-10K annotations. They are dynamically generated from ReState's own outputs and maintained during training, exposing the state-update operator to model-induced reasoning states that more closely resemble those encountered during recursive inference. Therefore, RAD provides diverse pre-constructed reasoning-state supervision, while CRSL further introduces dynamically evolving historical states to bridge the training-inference reasoning-state gap.

Table 5: The Break-Down ablation of ReState on the M^{3}CoT dataset. SHS and ISR denote the training-time self-history state mechanism and inference-time iterative self-refinement in CRSL.

Scheme	URSM	CRSL		Qwen2.5-VL-7B	Qwen3-VL-8B	Qwen3.5-9B
		SHS	ISR			
Baseline	✗	✗	✗	79.42 (+5.57)	81.62 (+6.34)	84.73 (+4.27)
Baseline w/ URSM	✓	✗	✗	84.12 (+0.87)	86.58 (+1.38)	88.65 (+0.35)
Baseline w/ URSM, CRSL(SHS)	✓	✓	✗	84.17 (+0.82)	86.67 (+1.29)	88.83 (+0.17)
Baseline w/ URSM, CRSL(ISR)	✓	✗	✓	84.43 (+0.56)	87.14 (+0.82)	88.65 (+0.35)
ReState (Final)	✓	✓	✓	**84.99**	**87.96**	**89.00**

Table 6: The Break-Down ablation of ReState on the ScienceQA dataset. SHS and ISR denote the training-time self-history state mechanism and inference-time iterative self-refinement in CRSL.

Scheme	URSM	CRSL		Qwen2.5-VL-7B	Qwen3-VL-8B	Qwen3.5-9B
		SHS	ISR			
Baseline	✗	✗	✗	96.11 (+1.32)	96.89 (+1.27)	95.59 (+2.78)
Baseline w/ URSM	✓	✗	✗	97.31 (+0.12)	97.95 (+0.21)	98.04 (+0.33)
Baseline w/ URSM, CRSL(SHS)	✓	✓	✗	97.38 (+0.05)	98.07 (+0.09)	98.37 (+0.00)
Baseline w/ URSM, CRSL(ISR)	✓	✗	✓	97.31 (+0.12)	97.95 (+0.21)	98.07 (+0.30)
ReState (Final)	✓	✓	✓	**97.43**	**98.16**	**98.37**

B Data Statistics for VersaCoT-10K

VersaCoT-10K contains 10,000 high-quality single-image multiple-choice visual reasoning samples spanning eight heterogeneous visual reasoning domains. From Fig. 7, we characterize the dataset from three perspectives: domain diversity, reasoning-length distribution, and candidate-space variability.

1) Diverse Reasoning Domains. VersaCoT-10K covers eight visual reasoning domains derived from PMC-VQA (Zhang et al., 2023b), DrivingVQA (Corbière et al., 2026), Visual7W (Zhu et al., 2016), AI2D (Kembhavi et al., 2016), CV-Bench (Tong et al., 2024), GeoQA (Chen et al., 2021), ChartQA (Masry et al., 2022), and MMMU (Yue et al., 2024), respectively. These domains encompass medical reasoning, driving and safety, visual commonsense, scientific diagram, spatial perception, geometric and mathematical, chart computation, and multidisciplinary knowledge. Each domain contains 1,250 samples, of which 20% are reserved for testing. This balanced composition prevents the overall benchmark performance from being disproportionately dominated by any single domain and enables more comprehensive evaluation across heterogeneous reasoning patterns.

2) Reasoning-Length Distribution. The CoT annotations have an average length of 170.9 words and a median length of 162 words, with lengths ranging from 50 to 537 words. The detailed distributions across the eight domains are shown in Fig. 7. Although reasoning length alone does not directly reflect reasoning difficulty, the retained CoTs have undergone strict quality control for answer consistency, visual grounding, reasoning correctness, completeness, and relevance. Consequently, they provide explicit and informative supervision involving visual evidence interpretation, knowledge integration, relational reasoning, and intermediate computation across diverse tasks, supporting both multi-domain reasoning and reasoning-state refinement.

3) Variable Candidate Spaces. All samples are standardized into a closed-choice multiple-choice format while preserving variations in the number of candidate options inherited from different task settings. The candidate-set size ranges from 2 to 9, with four-option questions accounting for the largest proportion. Rather than enforcing a fixed answer space, this design preserves task-specific decision structures within a unified evaluation interface, reducing dependence on a single option configuration and providing more diverse multiple-choice reasoning conditions.

C More Detailed Experimental Setup

1) Datasets. To evaluate the effectiveness and generalizability of ReState, we conduct experiments on the proposed VersaCoT-10K dataset and two widely used multimodal reasoning benchmarks, ScienceQA (Lu et al., 2022) and M^{3}CoT (Chen et al., 2024). ScienceQA focuses on scientific reasoning and contains 21,208 samples, while M^{3}CoT contains 11,289 samples spanning science, commonsense, and mathematics. For ScienceQA and M^{3}CoT, we follow their original data splits to ensure fair comparison with prior work. For VersaCoT-10K, we use the training, validation, and test

Table 7: Ablation of different reasoning states in ReState on VersaCoT-10K. ϕ , $\mathcal{R}^+$, $\mathcal{R}^-$, and $\mathcal{R}^h$ denote empty, positive, negative, and self-generated historical reasoning states, respectively.

Scheme	ϕ	$\mathcal{R}^+$	$\mathcal{R}^-$	$\mathcal{R}^h$	Qwen2.5-VL-7B	Qwen3-VL-8B	Qwen3.5-9B
ReState (Final) w/o ϕ	✗	✓	✓	✓	75.70 (+1.35)	77.95 (+1.30)	80.15 (+1.75)
ReState (Final) w/o $\mathcal{R}^+$	✓	✗	✓	✓	75.65 (+1.40)	78.50 (+0.75)	79.50 (+2.40)
ReState (Final) w/o $\mathcal{R}^-$	✓	✓	✗	✓	76.45 (+0.60)	78.90 (+0.35)	81.75 (+0.15)
ReState (Final) w/o $\mathcal{R}^h$	✓	✓	✓	✗	76.20 (+0.85)	78.65 (+0.60)	79.45 (+2.45)
ReState (Final)	✓	✓	✓	✓	**77.05**	**79.25**	**81.90**

splits described in Section A. Following (Yu et al., 2026), we adopt the corresponding RAD versions of ScienceQA and M^{3}CoT, which provide positive and negative reasoning states for learning unified reasoning-state transitions.

2) Implementation Details. We conduct experiments with three MLLM backbones: Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), and Qwen3.5-9B (Qwen Team, 2026a). All models are trained for 5 epochs with a learning rate of 1×10^{-5}. The batch size is set to 4 for Qwen2.5-VL-7B and Qwen3-VL-8B, and to 2 for Qwen3.5-9B. The maximum input length is uniformly set to 4096. Following prior experimental settings (Yu et al., 2026) and accounting for differences in reasoning complexity and output length across datasets, the maximum output length is set to 512 for ScienceQA and M^{3}CoT and to 1024 for VersaCoT-10K. Image captions are generated using Qwen2.5-VL-72B (Bai et al., 2025b), following (Yu et al., 2026). During ReState training, URSM constructs two views for each sample in every epoch: one with the empty reasoning state ϕ and one with a non-empty reasoning state. For the non-empty view, positive and negative reasoning states are sampled with equal probability. Therefore, each sample contributes two training instances per epoch. Building on URSM, CRSL further maintains a sample-specific historical reasoning-state pool with a maximum capacity of 2, from which model-generated historical states can be reused as subsequent training inputs. At inference, Restate evaluates a fixed refinement trajectory consisting of 9 rounds for all backbones and datasets. Following the same protocol, we report the peak dataset-level accuracy along this trajectory to characterize the refinement behavior, while the complete round-wise results are provided in Fig. 4. This evaluation summarizes the refinement trajectory and does not involve per-sample iteration selection. For settings without iterative self-refinement, the model performs only one additional answer-generation or reasoning-correction round after the initial CoT generation, providing a comparable non-iterative inference setting. For Direct SFT, we use "Answer + CoT" as the supervision target. All experiments are conducted on 8 NVIDIA H20 GPUs.

D More Ablation Experiments

D.1 Break-Down Ablation on M^{3}CoT and ScienceQA

To further examine the contributions of individual components across different reasoning benchmarks, we conduct the same break-down ablations on M^{3}CoT and ScienceQA as those reported for VersaCoT-10K in the main text. Specifically, we evaluate URSM together with the two components of CRSL: training-time Self-History State (SHS) learning and inference-time Iterative Self-Refinement (ISR). The results are reported in Tables 5 and 6.

As shown in Table 5, ReState yields substantial and consistent gains on the more challenging M^{3}CoT benchmark. Introducing URSM alone improves Qwen2.5-VL-7B by 4.70% (from 79.42% to 84.12%), Qwen3-VL-8B by 4.96% (from 81.62% to 86.58%), and Qwen3.5-9B by 3.92% (from 84.73% to 88.65%). These improvements demonstrate the effectiveness of unifying reasoning construction, correct reasoning preservation, and erroneous reasoning correction as transitions within a shared reasoning-state space. Building on URSM, either SHS or ISR provides further improvements, indicating that both training-time exposure to model-induced historical states and inference-time recursive state feedback contribute additional benefits. When both are enabled, the full ReState reaches 84.99%, 87.96%, and 89.00% on the three backbones, corresponding to gains of 5.57%, 6.34%, and 4.27% over their respective baselines. The full model achieves the best performance across all three backbones, showing that unified state modeling, historical-state learning, and iterative self-refinement provide complementary benefits.

Table 8: Main results (%) on the VersaCoT-10K dataset. Learning = Learning and training methods. Size = backbone model size. MED = medical, DRIVE = driving and safety, COMMON = visual commonsense, DIAG = scientific diagram, SPAT = spatial and perceptual, GEO = geometric and mathematical, CHART = chart and quantitative, and MULTI = multidisciplinary. Red denotes the best result, and blue denotes the second best result.

Model	Learning	Size	MED	DRIVE	COMMON	DIAG	SPAT	GEO	CHART	MULTI	**Avg**
Random	-	-	22.00	46.80	26.80	26.27	39.87	22.93	22.00	25.60	29.03
Human	-	-	66.00	89.20	90.40	88.00	92.00	94.40	96.00	85.20	87.65
Open-source Models + Zero-shot											
Qwen2.5-VL-7B	*Zero-shot*	7B	48.00	52.00	88.80	72.40	76.80	70.40	82.40	48.80	67.45
Qwen3-VL-8B	*Zero-shot*	8B	51.60	57.20	88.80	79.20	89.20	68.80	83.60	54.00	71.55
Qwen3-VL-32B	*Zero-shot*	32B	60.00	51.20	87.20	81.20	77.60	80.40	92.80	53.60	73.00
Qwen3-VL-235B-A22B	*Zero-shot*	235B	63.20	66.80	88.40	87.60	90.80	91.20	94.00	63.20	80.65
Qwen3.5-9B	*Zero-shot*	9B	61.60	59.60	87.60	81.20	83.60	70.80	84.80	54.00	72.90
Qwen3.5-27B	*Zero-shot*	27B	67.60	62.80	91.20	90.40	86.80	80.00	95.20	62.00	79.50
Qwen3.5-35B-A3B	*Zero-shot*	35B	66.80	62.40	89.60	86.80	84.00	74.00	92.80	59.60	77.00
Qwen3.5-122B-A10B	*Zero-shot*	122B	66.40	62.80	89.20	90.80	87.60	86.80	93.20	63.20	80.00
Qwen3.5-397B-A17B	*Zero-shot*	397B	66.80	62.40	90.80	89.60	88.80	93.60	95.20	65.60	81.60
Qwen3.6-27B	*Zero-shot*	27B	68.80	59.60	90.00	88.40	86.40	87.60	93.60	62.00	79.55
Qwen3.6-35B-A3B	*Zero-shot*	35B	64.00	63.20	89.20	85.20	85.20	78.00	94.40	62.00	77.65
MiniMax-M3	*Zero-shot*	428B	60.80	58.00	88.00	80.80	79.20	73.60	92.00	61.60	74.25
GLM-4.6V	*Zero-shot*	106B	60.40	59.20	90.00	85.20	87.60	78.40	93.60	60.80	76.90
Kimi-K2.6	*Zero-shot*	1T	68.40	63.20	91.60	77.20	85.20	55.20	94.00	58.00	74.10
Kimi-K3	*Zero-shot*	2.8T	71.60	70.40	90.00	92.40	90.80	91.60	98.00	80.40	85.65
Close-source Models + Zero-shot											
Qwen-VL-Plus	*Zero-shot*	-	60.00	53.60	87.60	83.60	80.40	84.80	85.20	53.20	73.55
Qwen3-VL-Flash	*Zero-shot*	-	61.20	58.80	86.80	86.40	88.00	82.00	88.00	60.00	76.40
Qwen3-VL-Plus	*Zero-shot*	-	64.00	63.20	90.40	87.20	88.00	90.00	94.40	63.20	80.05
Qwen3.5-Flash	*Zero-shot*	-	65.60	62.00	90.00	89.20	84.40	73.60	92.40	61.20	77.30
Qwen3.5-Plus	*Zero-shot*	-	68.00	60.40	90.00	90.00	88.40	93.60	94.40	64.80	81.20
Qwen3.6-Flash	*Zero-shot*	-	62.80	62.00	89.20	87.60	85.60	80.00	91.60	58.40	77.15
Qwen3.6-Plus	*Zero-shot*	-	69.60	67.20	89.60	91.20	88.80	94.00	94.80	67.60	82.85
Qwen3.7-Flash	*Zero-shot*	-	65.60	62.80	84.00	86.80	84.80	74.40	90.00	60.00	76.05
Qwen3.7-Plus	*Zero-shot*	-	66.80	67.60	90.40	90.40	88.40	95.20	94.80	66.80	82.55
Qwen3.8-Max	*Zero-shot*	-	74.00	70.00	92.40	91.60	92.00	94.80	96.80	71.20	85.35
Grok-4.3	*Zero-shot*	-	61.60	66.00	87.20	83.60	80.00	55.20	93.60	63.20	73.80
Grok-4.5	*Zero-shot*	-	70.80	70.80	89.20	92.00	85.20	94.80	98.00	86.80	85.95
Grok-4.6	*Zero-shot*	-	74.00	71.60	88.00	92.80	87.20	95.20	98.00	88.00	86.85
GLM-5V-Turbo	*Zero-shot*	-	66.00	63.60	89.20	86.00	91.20	64.00	94.80	59.60	76.80
GPT-4o	*Zero-shot*	-	62.40	67.60	90.00	81.20	80.40	44.80	89.60	57.60	71.70
GPT-5.4	*Zero-shot*	-	67.60	72.00	89.60	85.20	82.40	65.60	95.20	63.60	77.65
GPT-5.5	*Zero-shot*	-	72.40	64.00	88.40	89.20	85.20	82.80	96.00	70.40	81.05
GPT-5.6-Luna	*Zero-shot*	-	62.80	58.00	88.80	83.20	77.20	49.20	90.40	58.40	71.00
GPT-5.6-Terra	*Zero-shot*	-	66.00	62.40	88.80	86.80	84.80	73.60	95.20	67.60	78.15
GPT-5.6-Sol	*Zero-shot*	-	70.80	69.20	90.00	92.00	82.80	87.60	97.20	74.40	83.00
Claude Haiku 4.5	*Zero-shot*	-	57.20	58.80	86.00	79.60	66.40	62.80	90.00	60.00	70.10
Claude Sonnet 4.6	*Zero-shot*	-	66.80	66.40	88.00	86.00	78.80	78.80	93.60	67.20	78.20
Claude Sonnet 5	*Zero-shot*	-	68.80	63.60	89.20	86.80	82.00	78.80	94.80	69.20	79.15
Claude Opus 4.8	*Zero-shot*	-	71.20	74.00	88.40	86.00	82.40	86.80	96.00	72.40	82.15
Claude Opus 5	*Zero-shot*	-	70.80	75.20	89.20	90.80	85.60	89.60	98.40	78.00	84.70
Seed-1.6	*Zero-shot*	230B	64.40	62.00	90.00	83.20	88.00	91.60	94.00	66.40	79.95
Seed-1.8	*Zero-shot*	-	64.80	58.80	90.80	88.00	85.60	91.60	94.80	66.40	80.10
Seed-2.0-Mini	*Zero-shot*	-	66.00	60.80	91.60	87.60	88.40	83.60	92.80	65.20	79.50
Seed-2.0-Pro	*Zero-shot*	-	70.80	66.40	88.80	91.60	88.00	95.60	96.00	72.40	83.70
Seed-2.1-Turbo	*Zero-shot*	-	72.80	67.60	88.40	92.80	87.20	96.00	96.40	72.00	84.15
Seed-2.1-Pro	*Zero-shot*	-	74.40	71.20	89.60	90.40	88.80	96.80	94.80	75.60	85.20
Fine-tuning Methods											
Qwen2.5-VL-7B	*Fine-tune*	7B	**61.20**	65.60	**90.00**	83.20	88.40	72.40	91.20	49.20	75.15
Qwen2.5-VL-7B + Multimodal-CoT	*Fine-tune*	7B	49.20	62.00	83.60	82.40	81.20	73.20	92.40	**57.20**	72.65
Qwen2.5-VL-7B + MIND	*Fine-tune*	7B	55.20	66.00	88.80	85.20	85.60	74.00	**94.80**	56.40	75.75
Qwen2.5-VL-7B + ReState (Ours)	*Fine-tune*	7B	57.20	**68.00**	88.40	**87.60**	**89.60**	**76.80**	94.00	54.80	**77.05**
Qwen3-VL-8B	*Fine-tune*	8B	**59.20**	60.80	88.40	87.60	90.40	76.00	90.80	52.00	75.65
Qwen3-VL-8B + Multimodal-CoT	*Fine-tune*	8B	50.00	66.00	83.20	82.00	90.80	83.20	**95.60**	**61.20**	76.50
Qwen3-VL-8B + MIND	*Fine-tune*	8B	54.80	67.60	**90.80**	84.80	90.00	**83.60**	94.40	60.80	78.35
Qwen3-VL-8B + ReState (Ours)	*Fine-tune*	8B	58.80	**69.20**	87.20	**89.60**	**91.60**	82.80	94.40	60.40	**79.25**
Qwen3.5-9B	*Fine-tune*	9B	61.20	70.00	86.80	90.00	90.00	74.00	94.40	56.00	77.80
Qwen3.5-9B + Multimodal-CoT	*Fine-tune*	9B	56.40	63.60	83.60	87.20	88.00	**88.80**	94.00	56.80	77.30
Qwen3.5-9B + MIND	*Fine-tune*	9B	63.20	68.80	86.40	**91.60**	90.80	84.00	**96.00**	60.80	80.20
Qwen3.5-9B + ReState (Ours)	*Fine-tune*	9B	**66.40**	**76.00**	**89.20**	91.20	**92.00**	83.60	95.20	**61.60**	**81.90**

A similar trend is observed on ScienceQA in Table 6. URSM alone improves the three backbones by 1.20% (from 96.11% to 97.31%), 1.06% (from 96.89% to 97.95%), and 2.45% (from 95.59% to 98.04%), respectively. Incorporating SHS further increases performance to 97.38%, 98.07%, and 98.37%. The full ReState achieves 97.43%, 98.16%, and 98.37%, improving over the corresponding baselines by 1.32%, 1.27%, and 2.78%. Compared with M^{3}CoT, the additional gains from SHS and ISR are smaller on ScienceQA, which is consistent with the relatively limited headroom under

already strong baseline performance. Nevertheless, the full ReState achieves the best result across all three backbones, indicating that reasoning-state modeling and refinement remain effective even when overall benchmark accuracy is already high.

Overall, the results on M^3CoT and ScienceQA are consistent with the observations on VersaCoT-10K in the main text. URSM provides the shared state-transition foundation for jointly modeling heterogeneous reasoning behaviors. SHS further exposes the state-update operator to model-induced historical states during training, reducing the training-inference reasoning-state gap, while ISR recursively feeds the current state back into the same operator to enable iterative refinement during inference. Their complementary gains across datasets and backbones demonstrate that the effectiveness of ReState is not specific to VersaCoT-10K, further supporting the general applicability of the proposed state-centric reasoning framework.

D.2 Effect of Learning with Different Reasoning States

To further investigate the roles of different reasoning states in ReState, we conduct state-level ablations on VersaCoT-10K using all three backbones. Starting from the full ReState, we separately remove the empty state ϕ, positive states $\mathcal{R}^+$, negative states $\mathcal{R}^-$, and historical states $\mathcal{R}^h$. As shown in Table 7, removing any state type degrades performance, whereas jointly using all four types consistently yields the strongest overall results. This demonstrates that the different reasoning states provide complementary supervision for unified reasoning-state learning.

Among the three pre-constructed states used by URSM, removing the empty state ϕ reduces accuracy by 1.35%, 1.30%, and 1.75% on Qwen2.5-VL-7B, Qwen3-VL-8B, and Qwen3.5-9B, respectively. This confirms the importance of explicitly learning to construct an initial reasoning state from an empty input state. Removing positive states $\mathcal{R}^+$ causes larger drops of 0.75%-2.40%, with Qwen3.5-9B decreasing from 81.90% to 79.50%. These results show that exposure to diverse valid reasoning states is important for learning to preserve correct evidence while further improving reasoning completeness and consistency. Removing negative states $\mathcal{R}^-$ results in smaller but consistent decreases of 0.15%-0.60% across the three backbones, indicating that erroneous states provide useful supervision for identifying and correcting reasoning errors in existing states.

Historical states $\mathcal{R}^h$ play a different role from the three pre-constructed URSM states. Removing them reduces performance by 0.60%–2.45%, with Qwen3.5-9B decreasing from 81.90% to 79.45%. Unlike ϕ, $\mathcal{R}^+$, and $\mathcal{R}^-$, which are predefined before training, historical states are dynamically generated by ReState itself and therefore reflect model-induced reasoning patterns and errors. Exposing the state-update operator to these states during training better aligns its input distribution with the states encountered during recursive inference. The observed degradation without $\mathcal{R}^h$ therefore further validates the role of SHS in mitigating the training-inference reasoning-state gap.

Overall, the four state types provide complementary supervision for ReState. The empty state ϕ supports reasoning construction, positive states $\mathcal{R}^+$ support correct reasoning preservation and refinement, and negative states $\mathcal{R}^-$ support erroneous reasoning correction. Historical states $\mathcal{R}^h$ further extend this predefined state space with dynamically generated model-induced states, enabling closed-loop reasoning-state learning. Although the magnitude of each contribution varies across backbones, removing any state type consistently reduces performance. These results jointly validate URSM's unified modeling of heterogeneous reasoning behaviors and CRSL's complementary modeling of historical state distributions, further supporting the proposed state-centric reasoning paradigm.

E Comprehensive Comparison with Existing Methods

E.1 Evaluation on the VersaCoT-10K Dataset

To provide a more comprehensive evaluation of multi-domain visual reasoning, we extend the open-source and closed-source model comparisons reported in Table 1 of the main text. The complete results are presented in Table 8. Overall, substantial performance differences are observed across model families, model scales, and capability configurations, demonstrating that VersaCoT-10K can effectively differentiate the cross-domain visual reasoning capabilities of current MLLMs.

Table 9: Main results (%) on the ScienceQA dataset. Learning = Learning and training methods. Size=backbone model size. NAT = natural science, SOC= social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. Red denotes the best result, and blue denotes the second best result.

Model	Learning	Size	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	**Avg**
Random	-	-	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67	39.83
Human	-	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
MCAN	*Fine-tune*	95M	56.08	46.23	58.09	59.43	51.17	55.40	51.65	59.72	54.54
Top-Down	*Fine-tune*	70M	59.50	54.33	61.82	62.90	54.88	59.79	57.27	62.16	59.02
BAN	*Fine-tune*	112M	60.88	46.57	66.64	62.61	52.60	65.51	56.83	63.94	59.37
DFAF	*Fine-tune*	74M	64.03	48.82	63.55	65.88	54.49	64.11	57.12	67.17	60.72
ViLT	*Fine-tune*	113M	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90	61.14
Patch-TRM	*Fine-tune*	90M	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50	61.42
VisualBERT	*Fine-tune*	111M	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92	61.87
GPT-3.5	*Few-shot*	175B	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/CoT	*Few-shot*	175B	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
ChatGPT w/CoT	*Few-shot*	175B	78.82	70.98	83.18	77.37	67.92	86.13	80.72	74.03	78.31
GPT-4 w/CoT	*Few-shot*	-	85.48	72.44	90.27	82.65	71.49	92.89	86.66	79.04	83.99
Chameleon (ChatGPT)	*Few-shot*	-	81.62	70.64	84.00	79.77	70.80	86.62	81.86	76.53	79.93
Chameleon (GPT-4)	*Few-shot*	-	89.83	74.13	89.82	88.27	77.64	92.13	88.03	83.72	86.54
Gemini+RMR	*Few-shot*	-	91.79	94.26	89.64	91.40	89.69	91.01	92.84	89.78	91.75
Qwen2.5-VL-7B	*Zero-shot*	7B	83.79	88.30	82.45	83.63	84.38	81.81	86.67	80.29	84.39
Qwen3-VL-8B	*Zero-shot*	8B	90.54	94.49	87.64	92.18	91.67	86.20	90.64	90.57	90.62
Qwen3.5-9B	*Zero-shot*	9B	91.34	94.38	88.55	92.72	92.81	86.69	92.47	89.06	91.25
Qwen3.5-Flash	*Zero-shot*	-	94.98	98.43	89.18	95.60	95.79	89.83	94.82	93.08	94.20
Qwen3.5-Plus	*Zero-shot*	-	97.20	97.86	95.73	97.12	96.88	95.68	97.17	96.57	96.96
Qwen3.6-Flash	*Zero-shot*	-	94.85	97.98	90.36	95.41	95.09	90.80	94.68	93.74	94.34
Qwen3.6-Plus	*Zero-shot*	-	96.98	97.19	94.91	96.97	96.58	94.77	96.77	95.98	96.49
Qwen3.7-Flash	*Zero-shot*	-	95.20	98.31	89.09	95.80	95.64	89.97	95.30	92.42	94.27
Qwen3.7-Plus	*Zero-shot*	-	96.89	97.98	94.91	97.12	96.93	94.77	96.95	95.98	96.60
Qwen3.8-Max	*Zero-shot*	-	98.31	99.55	96.55	98.68	98.81	96.38	98.31	97.76	98.11
MiniMax-M3	*Zero-shot*	-	93.16	95.05	88.91	93.45	92.96	89.34	93.61	90.38	92.45
Grok-4.6	*Zero-shot*	-	95.60	95.39	94.64	96.63	94.89	93.80	96.15	93.80	95.31
GLM-4.6V	*Zero-shot*	-	92.41	93.25	93.73	91.84	90.43	93.45	93.50	91.89	92.93
GPT-4o	*Zero-shot*	-	91.47	91.34	92.00	89.44	87.95	92.68	93.17	88.73	91.58
GPT-5.6-Luna	*Zero-shot*	-	92.14	88.41	91.09	90.57	88.35	91.08	92.91	87.80	91.09
GPT-5.6-Terra	*Zero-shot*	-	94.45	92.13	93.91	94.18	92.02	93.59	94.82	92.02	93.82
GPT-5.6-Sol	*Zero-shot*	-	98.18	95.95	98.82	97.65	96.58	98.75	98.38	96.97	97.88
Gemini-3-Flash	*Zero-shot*	-	98.31	98.20	97.91	98.04	97.97	97.84	98.46	97.69	98.18
Gemini-3.5-Flash	*Zero-shot*	-	98.52	98.43	97.64	98.19	98.02	97.98	98.57	97.82	98.30
Claude Haiku 4.5	*Zero-shot*	-	89.17	89.31	92.45	87.83	85.37	92.54	92.14	86.29	90.05
Claude Sonnet 5	*Zero-shot*	-	95.60	95.39	95.45	94.13	93.90	95.89	96.26	94.20	95.52
Claude Opus 5	*Zero-shot*	-	98.45	96.85	98.64	97.41	97.22	98.68	98.60	97.36	98.16
Seed-2.1-Turbo	*Zero-shot*	-	96.67	97.08	96.36	96.68	95.79	96.45	96.92	96.24	96.68
Seed-2.1-Pro	*Zero-shot*	-	97.11	97.98	96.27	97.41	96.68	96.24	97.10	97.03	97.08
Kimi-K3	*Zero-shot*	2.8T	97.96	98.31	99.09	98.00	97.57	98.82	98.46	98.09	98.33
LLaMA-Adapter	*Fine-tune*	6B	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
LLaVA	*Fine-tune*	13B	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LaVIN	*Fine-tune*	7B	89.25	94.94	85.24	88.51	87.46	88.08	90.16	88.07	89.41
UnifiedQA$_{base}$	*Fine-tune*	223M	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00	70.12
UnifiedQA$_{base}$ w/CoT	*Fine-tune*	223M	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82	74.11
Multimodal-CoT$_{base}$	*Fine-tune*	223M	84.06	92.35	82.18	82.75	82.75	84.74	85.79	84.44	85.31
DDCoT$_{base}$	*Fine-tune*	223M	88.72	86.84	84.91	87.59	83.34	88.08	88.58	85.10	87.34
MC-CoT$_{base}$	*Fine-tune*	223M	91.87	84.59	93.00	92.28	88.30	92.75	90.64	90.64	90.64
DPMM-CoT$_{base}$	*Fine-tune*	223M+83M	92.72	87.85	89.91	92.72	90.48	91.29	91.45	90.11	90.97
Multimodal-T-SciQ$_{base}$	*Fine-tune*	223M	91.52	91.45	92.45	91.94	90.33	92.26	92.11	91.10	91.75
MIND$_{base}$	*Fine-tune*	223M	93.07	96.74	87.09	92.42	92.76	89.27	92.91	91.17	92.29
Qwen2.5-VL-7B	*Fine-tune*	7B	97.02	97.86	92.82	96.77	96.33	94.15	96.92	94.66	96.11
Qwen2.5-VL-7B + Multimodal-CoT	*Fine-tune*	7B	97.47	97.19	90.82	97.46	96.58	92.40	96.04	95.06	95.68
Qwen2.5-VL-7B + MIND	*Fine-tune*	7B	97.56	98.20	93.36	97.56	97.03	94.49	97.06	95.78	96.60
Qwen2.5-VL-7B + ReState (Ours)	*Fine-tune*	7B	**97.91**	**99.44**	**94.82**	**98.04**	**98.17**	**95.54**	**97.91**	**96.57**	**97.43**
Qwen3-VL-8B	*Fine-tune*	8B	98.13	97.19	94.09	98.00	97.12	95.12	96.95	96.77	96.89
Qwen3-VL-8B + Multimodal-CoT	*Fine-tune*	8B	97.42	97.08	94.27	97.26	96.28	95.33	96.92	95.85	96.53
Qwen3-VL-8B + MIND	*Fine-tune*	8B	97.96	98.31	94.64	97.90	97.37	95.61	97.32	96.90	97.17
Qwen3-VL-8B + ReState (Ours)	*Fine-tune*	8B	**99.07**	**99.10**	**95.55**	**99.17**	**98.81**	**96.38**	**98.27**	**97.96**	**98.16**
Qwen3.5-9B	*Fine-tune*	9B	95.69	96.40	94.73	95.31	94.15	95.68	95.89	95.06	95.59
Qwen3.5-9B + Multimodal-CoT	*Fine-tune*	9B	96.40	96.63	92.73	96.29	95.49	94.01	95.78	94.99	95.50
Qwen3.5-9B + MIND	*Fine-tune*	9B	95.78	97.64	97.18	95.45	94.60	**97.63**	97.10	95.52	96.53
Qwen3.5-9B + ReState (Ours)	*Fine-tune*	9B	**98.71**	**98.76**	**97.36**	**98.78**	**98.46**	**97.63**	**98.60**	**97.96**	**98.37**

Among the evaluated open-source zero-shot models (Bai et al., 2025b;a; Qwen Team, 2026a;b;c; Lai et al., 2026; Team et al., 2026b; Moonshot AI, 2026; Kimi Team, 2026), performance varies substantially on VersaCoT-10K. Kimi-K3 (Kimi Team, 2026) achieves the highest average accuracy

of 85.65%, followed by Qwen3.5-397B-A17B (Qwen Team, 2026a) at 81.60%. Qwen3-VL-235B-A22B (Bai et al., 2025a), Qwen3.5-122B-A10B (Qwen Team, 2026a), and Qwen3.6-27B (Qwen Team, 2026b) achieve 80.65%, 80.00%, and 79.55%, respectively. Although stronger and larger models often obtain higher accuracy, performance is not monotonically determined by total parameter count, indicating that model architecture, active parameterization, and training also affect cross-domain reasoning capability. These results highlight the ability of VersaCoT-10K to distinguish models with different architectural and capability configurations.

Several closed-source models (Bai et al., 2023; 2025a; Qwen Team, 2026a;e;f;d; xAI, 2026a;b;c; Team et al., 2026a; OpenAI, 2024; 2026d;a;b;c; Anthropic, 2025; 2026c;d;a;b; ByteDance Seed, 2025; Seed, 2026a;b; ByteDance Seed, 2026) achieve strong overall performance. Grok-4.6 (xAI, 2026c), Qwen3.8-Max (Qwen Team, 2026d), and Seed-2.1-Pro (ByteDance Seed, 2026) reach 86.85%, 85.35%, and 85.20%, respectively. Clear differences are also observed among capability configurations within the same model family. For example, the Qwen Plus variants generally outperform their corresponding Flash variants. GPT-5.6-Luna, Terra, and Sol (OpenAI, 2026c) achieve average accuracies of 71.00%, 78.15%, and 83.00%, respectively. A similar progression is observed within the Claude family, from 70.10% with Haiku 4.5 (Anthropic, 2025) to 84.70% with Opus 5 (Anthropic, 2026b). These results show that VersaCoT-10K captures not only differences across model families but also capability differences among variants within the same family.

The fine-grained results across the eight reasoning domains further reveal distinct capability profiles. Stronger models generally achieve high accuracy on *COMMON* and *CHART*, whereas *MED*, *DRIVE*, *GEO*, and *MULTI* exhibit substantially larger performance variation across models. For example, accuracy on *GEO* ranges from 44.80% to above 96.80% among the evaluated models, while *MULTI* similarly exhibits a wide performance range. These domains require different combinations of visual perception, domain knowledge, spatial understanding, and quantitative reasoning. Therefore, VersaCoT-10K provides both an overall measure of multi-domain reasoning through average accuracy and a fine-grained view of domain-specific capability differences.

More importantly, under matched backbone and training settings, ReState consistently achieves the highest average performance across all three evaluated backbones. With Qwen2.5-VL-7B (Bai et al., 2025b), ReState reaches 77.05%, outperforming Direct SFT, Multimodal-CoT (Zhang et al., 2024), and MIND (Yu et al., 2026) by 1.90% (from 75.15% to 77.05%), 4.40% (from 72.65% to 77.05%), and 1.30% (from 75.75% to 77.05%), respectively. With Qwen3-VL-8B (Bai et al., 2025a), ReState achieves 79.25%, corresponding to improvements of 3.60% (from 75.65% to 79.25%), 2.75% (from 76.50% to 79.25%), and 0.90% (from 78.35% to 79.25%). With Qwen3.5-9B (Qwen Team, 2026a), ReState reaches 81.90%, improving by 4.10% (from 77.80% to 81.90%), 4.60% (from 77.30% to 81.90%), and 1.70% (from 80.20% to 81.90%), respectively.

The improvements of ReState are also not confined to a single reasoning domain. With Qwen3.5-9B, ReState achieves the best results on *MED*, *DRIVE*, *COMMON*, *SPAT*, and *MULTI*. It also improves multiple reasoning domains with Qwen2.5-VL-7B and Qwen3-VL-8B. Although other methods retain advantages in several individual domains, ReState consistently achieves the highest average accuracy across all three backbones. This suggests that its gains arise from broadly improved reasoning-state modeling rather than optimization toward a particular domain.

Notably, ReState with Qwen3.5-9B achieves 81.90%, surpassing all evaluated open-source zero-shot models in Table 8 except Kimi-K3 (Kimi Team, 2026). For example, Qwen3.5-397B-A17B (Qwen Team, 2026a) and Qwen3-VL-235B-A22B (Bai et al., 2025a) achieve 81.60% and 80.65%, respectively. We emphasize that fine-tuned and zero-shot models are evaluated under different settings and therefore do not constitute strictly matched comparisons. Nevertheless, these results show that ReState can substantially enhance the reasoning capability of a medium-scale MLLM, enabling competitive multi-domain performance relative to much larger models.

Overall, the extended results demonstrate two complementary properties. First, VersaCoT-10K provides a discriminative benchmark that reveals substantial performance differences across model families, capability configurations, and reasoning domains. Second, under controlled matched-backbone comparisons, ReState consistently improves multi-domain reasoning performance across different MLLMs. Together, these results further support the effectiveness and cross-model adaptability of the proposed state-centric reasoning framework.

Table 10: Main results (%) on the M^3CoT dataset. The "Random" and "Human" results are averaged over three trials. Red denotes the best result, and blue denotes the second best result.

Model	Size	Science			Commonsense			Mathematics			**Total**
		Lang	Natural	Social	Physical	Social	Temporal	Algebra	Geometry	Theory	
Random	-	32.70	30.62	26.71	32.97	22.22	20.33	35.71	27.50	23.81	28.56
Human	-	97.83	92.62	94.31	96.28	92.41	88.71	87.23	88.75	85.71	91.61
Tool-Usage Methods											
HuggingGPT	175B	17.57	20.93	10.33	8.70	14.75	9.76	11.35	22.50	9.52	14.60
VisualChatGPT	>175B	30.09	36.28	7.78	43.48	29.92	33.33	21.99	21.25	28.57	25.92
IdealGPT	-	31.73	31.63	26.23	56.52	50.00	26.83	20.57	30.00	38.10	32.19
Chameleon	-	43.87	26.05	25.44	39.13	37.30	48.78	17.73	26.25	23.81	34.29
Zero-shot Methods											
Kosmos-2	2B	10.43	28.61	21.18	33.33	17.77	28.46	21.43	21.25	14.29	23.17
InstructBLIP	13B	38.39	30.52	26.27	76.67	70.66	35.77	30.00	22.50	19.05	35.94
LLava-V1.5	13B	36.97	27.46	20.22	52.22	23.55	27.64	22.86	45.00	4.76	27.05
CogVLM	17B	52.61	37.42	26.91	55.56	54.13	29.27	29.29	32.50	23.81	37.19
Gemini	-	73.93	41.25	31.21	56.67	71.49	62.60	30.71	27.50	28.57	45.17
GPT4V	-	80.09	54.66	43.95	87.78	67.77	82.11	42.14	43.75	42.86	56.95
Qwen2.5-VL-7B	7B	83.89	59.77	47.45	81.11	76.03	79.67	41.43	30.00	28.57	59.79
Qwen3-VL-8B	8B	86.73	65.77	45.86	82.22	77.27	77.24	34.29	32.50	38.10	61.43
Qwen3.5-9B	9B	89.57	66.28	47.45	87.78	80.99	78.86	30.71	33.75	38.10	62.81
Qwen3.5-Flash	-	94.79	69.99	52.87	87.78	79.75	88.62	45.00	45.00	33.33	67.60
Qwen3.5-Plus	-	94.31	76.37	63.06	86.67	83.47	92.68	62.86	45.00	66.67	74.42
Qwen3.6-Flash	-	94.79	74.07	53.50	88.89	81.40	86.18	55.71	46.25	33.33	69.93
Qwen3.6-Plus	-	93.84	76.88	62.26	88.89	80.58	91.87	65.00	45.00	47.62	74.03
Qwen3.7-Flash	-	94.31	72.41	53.50	90.00	76.86	88.62	56.43	55.00	52.38	69.54
Qwen3.7-Plus	-	94.31	79.69	60.99	90.00	78.51	92.68	62.86	56.25	52.38	74.85
Qwen3.8-Max	-	94.31	83.27	67.68	90.00	82.23	94.31	61.43	60.00	57.14	78.43
MiniMax-M3	-	90.05	67.05	50.00	85.56	75.62	87.80	50.71	38.75	57.14	65.19
Grok-4.6	-	93.84	85.19	73.89	91.11	75.62	94.31	70.71	72.50	53.38	81.02
GLM-4.6V	-	92.89	62.84	43.95	86.67	76.86	82.93	31.43	33.75	42.86	60.83
GPT-4o	-	95.26	57.85	43.95	91.11	79.34	78.05	30.71	33.75	33.33	59.40
GPT-5.6-Luna	-	94.31	62.45	42.20	88.89	76.45	78.86	46.43	42.50	33.33	61.30
GPT-5.6-Terra	-	94.79	68.84	50.64	86.67	76.03	86.99	60.71	38.75	42.86	66.91
GPT-5.6-Sol	-	98.10	78.80	66.08	88.89	79.75	91.06	73.57	48.75	61.90	76.75
Gemini-3-Flash	-	99.05	80.72	71.82	91.11	83.33	94.31	75.00	61.25	71.43	80.33
Gemini-3.5-Flash	-	99.05	85.95	72.45	93.33	85.12	95.12	78.57	63.75	85.71	82.96
Claude Haiku 4.5	-	93.36	54.66	36.46	87.78	72.73	79.67	33.57	32.50	23.81	55.44
Claude Sonnet 5	-	95.73	76.37	57.01	87.78	79.75	90.24	59.29	37.50	52.38	71.83
Claude Opus 5	-	97.63	83.65	65.45	90.00	80.17	95.12	88.57	66.25	61.90	79.98
Seed-2.1-Turbo	-	96.68	78.80	64.65	88.89	80.99	91.06	65.71	58.75	52.38	76.14
Seed-2.1-Pro	-	96.68	72.67	65.13	88.89	78.93	89.43	68.57	56.25	28.57	73.77
Kimi-K3	2.8T	97.16	86.72	70.38	91.11	78.51	95.93	77.86	60.00	57.14	81.32
Fine-tuning Methods											
LLaMA-Adaper	7B	62.56	72.29	30.21	76.92	59.67	72.36	30.71	38.75	38.10	54.89
LLaVA-V1.5	13B	68.72	72.41	40.86	83.52	64.61	69.11	35.71	45.00	38.10	59.50
CogVLM	17B	65.88	77.52	29.09	81.32	65.43	75.61	35.71	46.25	47.62	58.25
Multimodal-CoT$_{base}$	223M	41.71	46.49	39.90	59.34	60.91	27.64	48.57	35.00	28.57	44.85
Multimodal-CoT$_{large}$	738M	45.50	50.19	43.56	63.74	64.61	33.33	40.71	61.25	28.57	48.73
MC-CoT$_{base}$	223M	53.55	63.98	43.56	61.54	69.55	29.27	42.86	33.75	28.57	53.51
MC-CoT$_{large}$	738M	42.65	67.43	50.56	58.24	60.49	56.10	57.86	62.50	14.29	57.69
MIND$_{base}$	223M	72.04	63.09	43.31	82.22	65.29	44.72	49.29	61.25	33.33	57.38
MIND$_{large}$	738M	79.62	66.41	48.57	81.11	69.42	51.22	50.71	61.25	47.62	61.56
Qwen2.5-VL-7B	7B	92.42	91.83	60.35	84.44	**80.17**	86.18	72.14	70.00	71.43	79.42
Qwen2.5-VL-7B + Multimodal-CoT	7B	87.20	89.27	61.62	83.33	74.38	86.18	56.43	61.25	47.62	76.32
Qwen2.5-VL-7B + MIND	7B	92.42	**97.32**	**76.27**	83.33	75.62	**88.62**	66.43	71.25	61.90	84.81
Qwen2.5-VL-7B + ReState (Ours)	7B	**94.79**	**97.32**	71.82	**85.56**	78.10	86.18	**75.71**	**78.75**	**76.19**	**84.99**
Qwen3-VL-8B	8B	93.84	90.55	68.31	**86.67**	80.99	89.43	74.29	67.50	66.67	81.62
Qwen3-VL-8B + Multimodal-CoT	8B	90.05	92.72	65.92	83.33	67.36	90.24	62.14	62.50	66.67	78.95
Qwen3-VL-8B + MIND	8B	94.79	97.96	**79.62**	85.56	80.17	**95.12**	68.57	**77.50**	66.67	87.45
Qwen3-VL-8B + ReState (Ours)	8B	**95.73**	**98.47**	78.03	83.33	**83.88**	92.68	**76.43**	76.25	**76.19**	**87.96**
Qwen3.5-9B	9B	93.36	96.42	74.52	85.56	80.58	84.55	70.00	71.25	61.90	84.73
Qwen3.5-9B + Multimodal-CoT	9B	91.00	89.66	59.87	85.56	67.77	82.93	60.00	70.00	33.33	75.93
Qwen3.5-9B + MIND	9B	95.26	**98.08**	**81.69**	85.56	81.40	85.37	72.86	**83.75**	57.14	88.09
Qwen3.5-9B + ReState (Ours)	9B	**95.73**	97.83	80.89	**88.89**	**83.47**	**86.99**	**80.00**	**83.75**	**90.48**	**89.00**

E.2 Evaluation on the ScienceQA Dataset

To provide a more comprehensive evaluation on a standard scientific reasoning benchmark, we extend the ScienceQA comparisons reported in Table 2 of the main text. The complete results in Table 9 cover traditional visual question answering methods (Yu et al., 2019; Anderson et al., 2018; Kim et al., 2018; Gao et al., 2019; Kim et al., 2021; Lu et al., 2021; Li et al., 2020), few-shot reasoning methods (Wei et al., 2022; Achiam et al., 2023; Team et al., 2023; Lu et al., 2023; Tan et al., 2024b), several families of zero-shot MLLMs (Bai et al., 2025b;a; Qwen Team, 2026a;e;f; Lai et al., 2026;

Team et al., 2026b; xAI, 2026c; Qwen Team, 2026d; OpenAI, 2024; 2026c; Google DeepMind, 2025; 2026; Anthropic, 2025; 2026d;b; ByteDance Seed, 2026; Kimi Team, 2026), and additional learning-based MCoT methods (Zhang et al., 2023a; Liu et al., 2024; Luo et al., 2023; Khashabi et al., 2020; Zhang et al., 2024; Zheng et al., 2023; Tan et al., 2024a; He et al., 2024; Wang et al., 2024a; Yu et al., 2026; Bai et al., 2025b;a; Qwen Team, 2026a). Overall, recent MLLMs already achieve high accuracy on ScienceQA, providing a complementary setting for evaluating whether ReState can still yield consistent gains when the remaining performance headroom is limited.

Across different method categories, ScienceQA performance improves substantially with advances in model capability and reasoning mechanisms. Traditional vision-language models, including MCAN (Yu et al., 2019), BAN (Kim et al., 2018), ViLT (Kim et al., 2021), and VisualBERT (Li et al., 2020), mainly achieve average accuracies in the range of 54.54%-61.87%. With stronger language models and explicit reasoning mechanisms, performance increases markedly. GPT-4 w/CoT (Achiam et al., 2023), Gemini+RMR (Team et al., 2023; Tan et al., 2024b), and MIND$_{base}$ (Yu et al., 2026) reach 83.99%, 91.75%, and 92.29%, respectively. More recent closed-source MLLMs further raise the performance level: Qwen3.8-Max (Qwen Team, 2026d), GPT-5.6-Sol (OpenAI, 2026c), Gemini-3.5-Flash (Google DeepMind, 2026), Claude Opus 5 (Anthropic, 2026b), and Kimi-K3 (Kimi Team, 2026) all achieve accuracies around 98%. These results indicate that ScienceQA has entered a high-accuracy regime for strong MLLMs, making further improvements increasingly challenging. It is worth mentioning that ReState with Qwen3.5-9B also achieves competitive performance against strong zero-shot MLLMs. Its 98.37% accuracy exceeds Qwen3.8-Max, GPT-5.6-Sol, Gemini-3-Flash, Gemini-3.5-Flash, and Claude Opus 5, and is comparable to Kimi-K3 at 98.33%.

More importantly, under matched backbone and training settings, ReState consistently achieves the highest average performance across all three evaluated backbones. With Qwen2.5-VL-7B (Bai et al., 2025b), ReState reaches 97.43%, outperforming Direct SFT, Multimodal-CoT (Zhang et al., 2024), and MIND (Yu et al., 2026) by 1.32% (from 96.11% to 97.43%), 1.75% (from 95.68% to 97.43%), and 0.83% (from 96.60% to 97.43%), respectively. With Qwen3-VL-8B (Bai et al., 2025a), ReState achieves 98.16%, corresponding to gains of 1.27% (from 96.89% to 98.16%), 1.63% (from 96.53% to 98.16%), and 0.99% (from 97.17% to 98.16%). With Qwen3.5-9B (Qwen Team, 2026a), ReState further reaches 98.37%, improving by 2.78% (from 95.59% to 98.37%), 2.87% (from 95.50% to 98.37%), and 1.84% (from 96.53% to 98.37%), respectively. These consistent improvements demonstrate that ReState remains effective across different MLLM backbones even when baseline performance is already high.

The fine-grained results further show that the gains of ReState are not concentrated in a particular ScienceQA subset. With Qwen3-VL-8B, ReState achieves the strongest matched-backbone results across all reported dimensions, while improving the average accuracy from 97.17% with MIND to 98.16%. Similar overall trends are observed with Qwen2.5-VL-7B and Qwen3.5-9B. These results indicate that the benefits of ReState extend across different subject categories, context conditions, and grade levels rather than being driven by a particular subset.

Overall, the extended ScienceQA results demonstrate that ReState consistently improves performance across MLLM backbones in a high-accuracy evaluation regime. Together with its broad gains across ScienceQA subsets, these results further support the effectiveness of unified reasoning-state modeling and closed-loop reasoning-state refinement, as well as the cross-model adaptability of ReState.

E.3 Evaluation on the M^{3}CoT Dataset

To provide a more comprehensive evaluation of complex multimodal reasoning, we extend the M^{3}CoT comparisons reported in Table 3 of the main text. The complete results in Table 10 cover tool-usage methods (Shen et al., 2023; Wu et al., 2023; You et al., 2023; Lu et al., 2023), several families of zero-shot MLLMs (Peng et al., 2023; Dai et al., 2023; Liu et al., 2024; Wang et al., 2024b; Team et al., 2023; OpenAI, 2023; Bai et al., 2025b;a; Qwen Team, 2026a;e;f; Lai et al., 2026; Team et al., 2026b; xAI, 2026c; Qwen Team, 2026d; OpenAI, 2024; 2026c; Google DeepMind, 2025; 2026; Anthropic, 2025; 2026d;b; ByteDance Seed, 2026; Kimi Team, 2026), and additional fine-tuning methods (Zhang et al., 2023a; Liu et al., 2024; Wang et al., 2024b; Zhang et al., 2024; Tan et al., 2024a; Yu et al., 2026; Bai et al., 2025b;a; Qwen Team, 2026a). M^{3}CoT spans science, commonsense, and mathematics, with subtasks characterized by substantially different knowledge structures and reasoning patterns, providing a challenging testbed for heterogeneous multimodal reasoning.

Question: Based on the figure, what is the solution to the small loan dilemma?

Options: **VersaCoT-10K #2171**

A. Appropriately reduce loan interest

B. Discover more small and medium-sized enterprise customers

C. Improve loan trust

D. Understand the problem of poverty as the inability to pay someone back

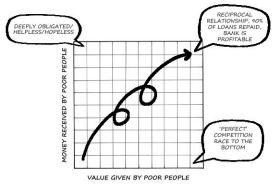


Figure 1.18 Why microfinance can work

Reference CoT:

Analyze the Graph's Annotations: The top-left speech bubble characterizes a state of being "DEEPLY OBLIGATED / HELPLESS / HOPELESS." This represents the traditional view or the starting condition of poverty where individuals are seen as unable to break the cycle of debt or lack the means to contribute value back. The bottom-right bubble mentions a "'PERFECT' COMPETITION RACE TO THE BOTTOM," indicating a failure of standard market mechanisms in this context. The top-right bubble, which is the destination of the arrow, describes a "RECIPROCAL RELATIONSHIP" where "90% OF LOANS REPAID" and the "BANK IS PROFITABLE." **Identify the Dilemma and Solution:** The "small loan dilemma" typically refers to the high risk and perceived impossibility of lending to the poor because they are viewed as having no collateral or means to repay. The graph shows that as the "VALUE GIVEN BY POOR PEOPLE" increases (moving right on the x-axis), the outcome shifts from "HELPLESS/HOPELESS" to a "RECIPROCAL RELATIONSHIP." The success of microfinance, as shown in the top-right bubble, is based on achieving a high repayment rate (90%). This directly addresses the core dilemma: the perceived inability of the poor to pay back loans. Therefore, the solution involves a conceptual shift: understanding that the fundamental problem of poverty in a lending context is the perceived or actual "inability to pay someone back." By framing the problem this way, microfinance can create a "reciprocal relationship" that enables repayment and profitability. **Evaluate the Choices:** (A), (B), and (C) are not explicitly supported by the specific labels in the diagram. (D) "Understand the problem of poverty as the inability to pay someone back" aligns with the transition from the "HELPLESS/HOPELESS" state to the "RECIPROCAL RELATIONSHIP" where repayment is finally achieved. This understanding allows for the design of a system that facilitates value exchange and high repayment rates. **Conclusion:** The graph demonstrates that by addressing the core issue of repayment through reciprocity, the dilemma of lending to the poor is solved.

True Answer: The answer is (D).

Direct SFT

Output CoT: ..... The visual path of the curve represents a transition from a state of financial distress and lack of trust to a state of successful repayment and mutual benefit. For microfinance to move from the first point to the second, borrowers must transition from being "deeply obligated" or "hopeless" to becoming reliable actors in a "reciprocal relationship." This shift requires overcoming the initial skepticism and lack of faith that poor people often have in financial institutions. ...While understanding poverty as an inability to pay (D) is a related concept, the diagram specifically highlights the need to build the trust necessary for a "reciprocal relationship" to flourish. Therefore, improving trust is the key solution represented by the positive trajectory of the curve.

Model Choice: C. Improve loan trust

Multimodal-CoT

Output CoT: ..... The curve rises as the value given by poor people increases, but then levels off, suggesting diminishing returns. ..."RECIPROCAL RELATIONSHIP, 90% OF LOANS REPAID, BANK IS PROFITABLE" suggests that microfinance works through a mutual, trusting lender-borrower relationship. ... By building a trust-based reciprocal relationship, microfinance becomes sustainable, with high repayment rates and bank profitability. Therefore, the solution is to improve trust between the lender and borrower. Comparing the choices:(A) Reducing interest rates does not address the relational aspect. (B) Finding more customers is not the core solution. (C) Improving loan trust matches the "RECIPROCAL RELATIONSHIP" shown in the graph. (D) Defining poverty as inability to pay is treated as a negative condition that microfinance overcomes. Thus, the graph points to trust as the core solution.

Model Choice: C. Improve loan trust

MIND

Output CoT: ..... Initial State: The curve starts steep, indicating poor people's minimal financial input yields minimal returns, reflecting "DEEPLY OBLIGATED / HELPLESS / HOPELESS." Microfinance Shift: The curve flattens, showing improved outcomes. Key Insight: The "RECIPROCAL RELATIONSHIP, 90% OF LOANS REPAID" suggests trust and mutual benefit. Final State: The curve reaches a point where effort equals reward, avoiding "PERFECT COMPETITION RACE TO THE BOTTOM." The solution is building trust to improve repayment rates, not reducing interest or expanding customer bases. Poverty isn't about repayment inability but the lack of financial tools.

Model Choice: C. Improve loan trust

ReState (Ours)

Output CoT: ..... The graph reveals that traditional lending fails in poverty because "value given" is so low that "money received" is insufficient, creating a debt cycle. Microfinance solves this by fostering a "reciprocal relationship" with high repayment rates. Option (D) correctly frames poverty as the inability to repay, which the graph addresses. Options (A), (B), and (C) are not the focus, as the diagram emphasizes the debt trap over interest rates, customer acquisition, or trust.

Model Choice: D. Understand the problem of poverty as the inability to pay someone back

Figure 8: Qualitative comparison on an example of small loan dilemma reasoning in VersaCoT-10K.

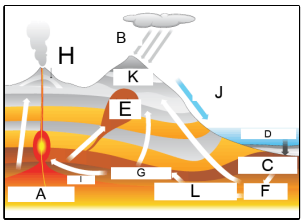
Existing methods exhibit substantial performance variation on M^3CoT. Early tool-usage methods achieve average accuracies below 35%, while GPT-4V (OpenAI, 2023) reaches 56.95%. With stronger MLLM backbones, zero-shot accuracy increases to 59.79%, 61.43%, and 62.81% for Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), and Qwen3.5-9B (Qwen Team, 2026a), respectively. More recent MLLMs achieve substantially stronger results: Gemini-3.5-Flash (Google DeepMind, 2026), Kimi-K3 (Kimi Team, 2026), Grok-4.6 (xAI, 2026c), and Gemini-3-Flash (Google DeepMind, 2025) reach 82.96%, 81.32%, 81.02%, and 80.33%, respectively. Despite these advances,

Question: The direct result of melting is represented at:

VersaCoT-10K #472

Options:

A. K | B. L | C. A | D. E



Reference CoT:
1. **Analyze the labels and arrows:** The diagram shows various processes like weathering (B), erosion/transport (J), sedimentation (D), and lithification (C/F).
2. **Locate the melting process:** Melting occurs deep underground where temperatures are high enough to liquefy rock. In the diagram, arrows from metamorphic rock regions (near L and I) point downward into a bright orange and yellow zone at the very bottom of the cross-section.
3. **Identify the product of melting:** The direct result of melting solid rock is magma. Label **A** is placed directly within this bright, molten-looking reservoir at the base of the diagram, which feeds the volcanic activity (H) and intrusive bodies (E).
4. **Evaluate other options:**
 * **K** represents rocks at the surface, which are the result of cooling or uplift, not melting.
 * **L** represents metamorphic rocks, which are formed by heat and pressure but remain in a solid state.
 * **E** represents an intrusive igneous rock body (pluton), which is the result of magma *cooling and solidifying*, not the melting process itself.

True Answer: The answer is (C).

Direct SFT

Output CoT: ..... evaporation (C) is water vapor rising from the ocean; and melting (E) occurs at the base of the mountain, where solid ice or snow changes into liquid water and then flows as runoff (J). The question asks for the direct result of melting. Melting converts solid ice or snow into liquid water. Label E depicts this transition point at the base of the mountain. Although J shows the resulting runoff, it is a downstream effect rather than the direct physical result of the phase change itself. Therefore, E is the most accurate label. .....

Model Choice: D. E

Multimodal-CoT

Output CoT: ..... Label H shows a volcanic eruption; B shows precipitation; K points to the crust or surface layer; E points to a large orange mass beneath the surface, which usually represents magma or molten rock; A is at the bottom and represents solid Earth or crust; L indicates material movement or a landslide. Melting changes solid material into liquid. In a volcanic context, melting occurs in the mantle or crust and produces magma. Therefore, E represents the molten state produced by melting, while A represents solid crust.

Model Choice: D. E

MIND

Output CoT: ..... Locate the melting process: Melting occurs when heat from below (like mantle convection) or surface water (like glacial melt) transforms solid rock into magma. The diagram shows a downward arrow from the surface (labeled 'B' for rain and 'J' for a river) hitting the crust, with arrows moving downward into the mantle, representing the heat and water that cause melting. Trace the arrows to the result:Following the arrows from the surface impact zone (B and J) downward, they point to location 'E'. ..... Label L points to a downward-pointing arrow, representing the process of sediment deposition or burial ..... Therefore, 'E' represents the result of melting in this diagram.

Model Choice: D. E

ReState (Ours)

Output CoT: ..... Label Interpretation: A points to a luminous, red-orange zone deep underground, representing magma, the liquid form of rock. K highlights a towering mountain, typically formed through compression and uplift. E shows a reddish-brown intrusion, indicating a solidified igneous rock. L denotes loose particles, representing sediment created by weathering and erosion. Process Mapping: The arrow labeled I travels from the crust's base to the magma chamber (A). This depicts the process of heat and pressure turning solid rock into liquid magma. .....

Model Choice: C. A

Figure 9: Qualitative comparison on an example of rock cycle reasoning in VersaCoT-10K.

substantial performance differences remain across individual subtasks, reflecting the heterogeneous reasoning requirements of M^3CoT.

More importantly, under matched backbone and training settings, ReState consistently achieves the highest average performance across all three backbones. With Qwen2.5-VL-7B, ReState reaches 84.99%, outperforming Direct SFT, Multimodal-CoT (Zhang et al., 2024), and MIND (Yu et al., 2026) by 5.57% (from 79.42% to 84.99%), 8.67% (from 76.32% to 84.99%), and 0.18% (from 84.81% to 84.99%), respectively. With Qwen3-VL-8B, ReState achieves 87.96%, corresponding to gains of 6.34% (from 81.62% to 87.96%), 9.01% (from 78.95% to 87.96%), and 0.51% (from 87.45% to 87.96%). With Qwen3.5-9B, ReState further reaches 89.00%, improving by 4.27% (from 84.73% to 89.00%), 13.07% (from 75.93% to 89.00%), and 0.91% (from 88.09% to 89.00%), respectively.

In particular, the gains over Multimodal-CoT range from 8.67% to 13.07%, substantially larger than those observed on ScienceQA and VersaCoT-10K. This suggests that explicitly modeling and recursively refining reasoning states is especially beneficial on M^3CoT, where diverse knowledge structures and reasoning patterns place greater demands on intermediate reasoning. ReState with

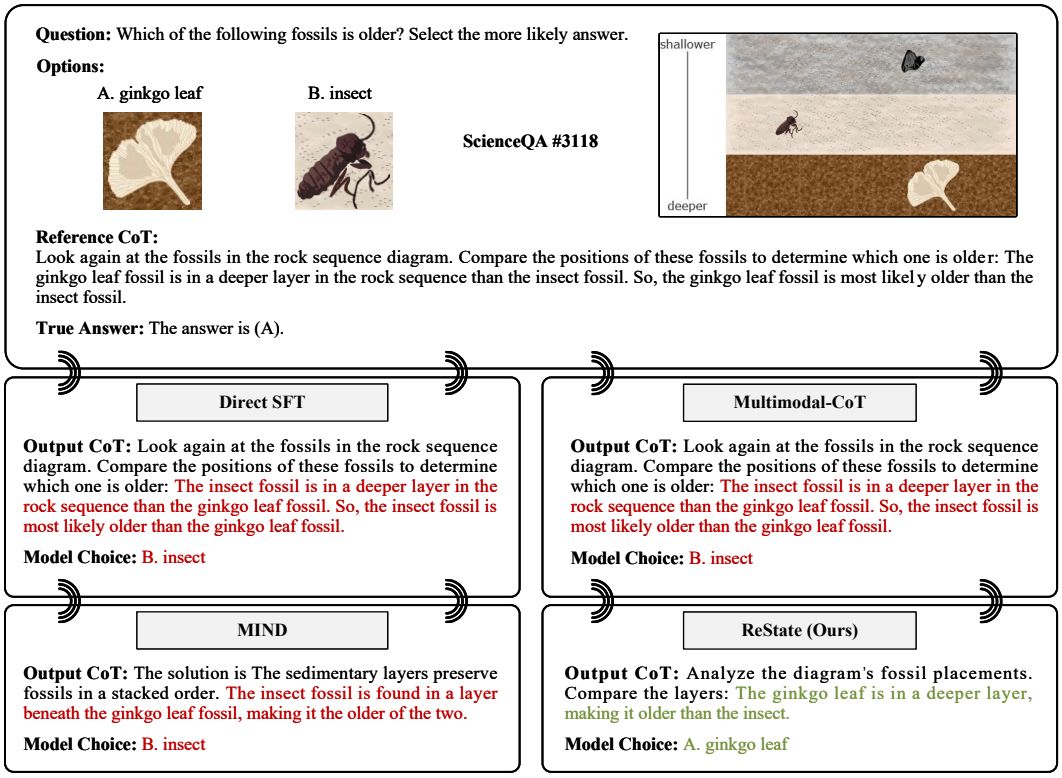


Figure 10: Qualitative comparison on an example of fossil age reasoning in ScienceQA.

Qwen3.5-9B also achieves an average accuracy of 89.00%, exceeding all evaluated zero-shot MLLMs in Table 10. We emphasize that fine-tuned and zero-shot models are evaluated under different settings and therefore do not constitute strictly matched comparisons. Nevertheless, this result shows that ReState can substantially enhance the reasoning capability of a medium-scale MLLM and enable highly competitive performance on complex multimodal reasoning. It also narrows the gap to the reported human performance of 91.61% to 2.61%.

The fine-grained results further show that the gains of ReState extend across science, commonsense, and mathematics. With Qwen3.5-9B, ReState achieves the best results within the matched-backbone comparison on multiple subtasks, including 80.00% on *Algebra*, 83.75% on *Geometry*, and 90.48% on *Theory*. In particular, *Theory* accuracy increases from 61.90% with Direct SFT to 90.48% with ReState. Similar improvements across multiple subtasks indicate that the overall performance gain is not driven by a single reasoning category, but instead reflects broader benefits across heterogeneous reasoning patterns.

Overall, the extended M^{3}CoT results demonstrate particularly clear gains from ReState on a benchmark with diverse and challenging reasoning requirements. ReState consistently improves average performance across all three backbones and remains effective across scientific, commonsense, and mathematical reasoning. Together with the results on VersaCoT-10K and ScienceQA, these findings further support the effectiveness and cross-dataset applicability of reformulating multimodal reasoning from stage-centric pipelines into state-centric iterative refinement.

F Qualitative Analysis

F.1 Comparison with Different Methods

To further examine the reasoning behaviors of different methods, we present representative examples from VersaCoT-10K, ScienceQA, and M^{3}CoT, comparing Direct SFT, Multimodal-CoT, MIND, and ReState in Figs. 8 to 13. Beyond final-answer correctness, these examples reveal clear differences in visual evidence interpretation, relation modeling, and reasoning consistency.

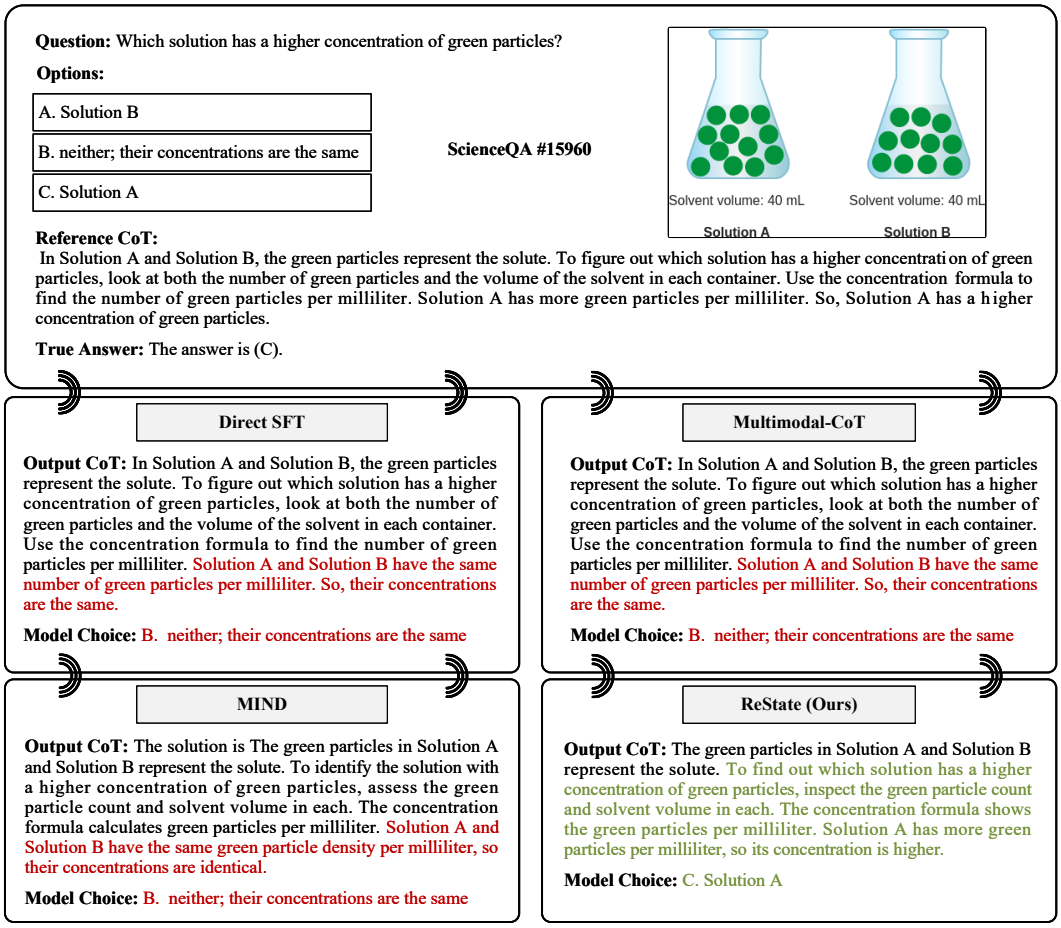


Figure 11: Qualitative comparison on an example of solution concentration reasoning in ScienceQA.

On VersaCoT-10K, Direct SFT, Multimodal-CoT, and MIND are more susceptible to misleading local visual or textual cues. In the "small loan dilemma" example, all three methods overemphasize the "reciprocal relationship" and high repayment rate, and consequently misidentify loan trust as the central issue. In contrast, ReState reasons over the broader semantic structure of the diagram and correctly identifies repayment ability as the key factor. Similarly, in the "rock cycle" example, the compared methods incorrectly identify E as the direct product of melting. ReState instead distinguishes the magma region from the solidified intrusive rock and correctly identifies A as the material produced by melting. These examples suggest that ReState more reliably captures global visual structures and process relations rather than relying primarily on local surface associations.

A similar pattern is observed on ScienceQA and M^3CoT. In the "fossil age" and "solution concentration" examples, the compared methods make incorrect predictions due to errors in spatial interpretation, quantitative comparison, or proportional reasoning, whereas ReState more accurately integrates the positions, quantities, and relations. In the "fine-grained species recognition" and "tangram recognition" examples, ReState also combines multiple visual attributes and global structural cues more effectively, reducing errors caused by over-reliance on isolated appearance features.

Overall, the compared methods can produce fluent and superficially plausible CoTs even when their intermediate reasoning is grounded in incorrect visual observations or misleading local associations. Such errors may then propagate through the reasoning chain and affect the final prediction. In contrast, ReState organizes and updates explicit reasoning states around task-relevant multimodal evidence, yielding stronger consistency among visual observations, intermediate reasoning, and final answers. These qualitative examples provide intuitive evidence that state-centric reasoning can better support robust multimodal reasoning than conventional stage-centric pipelines.

Question: Study the image given and determine the scientific name of the organism depicted.

Options: **M³CoT #10111**

A. Felis nigripes | B. Lynx rufus | C. Neofelis nebulosa









Reference CoT:
1. Observe the image and identify the key visual characteristics of the organism.
2. Analyze the body size and shape of the organism. It appears to have a medium-sized, compact body with short legs and a relatively long tail.
3. Examine the fur coloration of the organism. It displays a distinct coat color, which in this case is black or dark brown.
4. Evaluate the facial features of the organism. It has a rounded face with distinctive ear tufts.
5. Compare the visual traits observed to the provided options.
6. Based on the visual traits, determine that option A closely matches the described organism.
7. Understand that the scientific name "Felis nigripes" appropriately represents this species.
8. Recognize that the species described is commonly known as the black-footed cat.
9. For these reasons, the answer is option (A).

True Answer: The answer is (A).

Direct SFT

Output CoT: Step 1: Identify the key visual characteristics of the organism based on the image:
- Size and Shape: The organism is a medium-sized cat with a compact and muscular body.
- Fur Color and Pattern: The fur is predominantly reddish-brown with black spots and stripes on its body.
- Ear Tufts: The organism has distinct ear tufts on top of its ears.
Step 2: Analyze the options:
- Option B closely matches the described organism based on its size, shape, fur color and pattern, and ear tufts.
Step 3: Determine the scientific name:
- The scientific name "Lynx rufus" accurately represents this species of medium-sized cat with reddish-brown fur and distinct ear tufts.

Model Choice: B. Lynx rufus

Multimodal-CoT

Output CoT: 1. Size and Shape: The organism in the image is observed to have the characteristics of a medium-sized cat, possessing a compact and muscular body.
2. Fur Color and Pattern: The majority of the organism's fur is reddish-brown, accompanied by several black spots and stripes located throughout its body.
3. Ear Tufts: Notably, the organism exhibits distinct ear tufts positioned atop its ears.
Based on the analysis of these visual traits, it can be concluded that option B is the most suitable match for the described organism. The scientific name "Lynx rufus" accurately represents this particular species of medium-sized cat, wherein its physical attributes include reddish-brown fur and prominent ear tufts.

Model Choice: B. Lynx rufus

MIND

Output CoT: The solution is 1. **Step 1: Physical Attributes**
- The organism exhibits a medium-sized, muscular frame with short legs and a tail of moderate length.
2. **Step 2: Coat Details**
- The fur is primarily reddish-brown, with black spots and stripes scattered throughout.
3. **Step 3: Ear Features**
- The ears are large and adorned with conspicuous tufts.
4. **Step 4: Species Identification**
- These characteristics uniquely correspond to option B.
5. **Step 5: Taxonomic Name**
- "Lynx rufus" is the correct scientific name for this species, reflecting its reddish-brown fur.

Model Choice: B. Lynx rufus

ReState (Ours)

Output CoT: 1. Size and Shape: The organism is a mid-sized cat with a sturdy build and short legs.
2. Fur Coloration: Its fur is dark brown or black, with a sleek and shiny appearance.
3. Facial Features: The face is round, with a black nose and prominent ear tufts.
These features correspond to option A, "Felis nigripes," the scientific name for the black-footed cat. Hence, the correct answer is A.

Model Choice: A. Felis nigripes

Figure 12: Qualitative comparison on an example of species recognition in M³CoT.

F.2 Iterative Self-Refinement Capability of ReState

We further visualize the evolution of reasoning states during recursive inference to examine the iterative self-refinement capability of ReState. As shown in Figs. 14 to 16, when an early reasoning state contains an incorrect judgment, ReState can feed the current state back into the same state-update operator, re-examine task-relevant visual evidence and reasoning relations, and progressively revise the erroneous state.

For example, in the highway scene in Fig. 14, the initial reasoning state is distracted by irrelevant facility information and incorrectly predicts that food is unavailable. After recursive state feedback, the model shifts its attention to the fork-and-knife icon that is directly relevant to the question, revises the answer to "Yes," and preserves the corrected reasoning in subsequent iterations. In the "food web" example in Fig. 15, the initial state constructs an incorrect matter-transfer relation. After receiving its own reasoning state as input again, the model re-examines the arrow directions and inter-organism

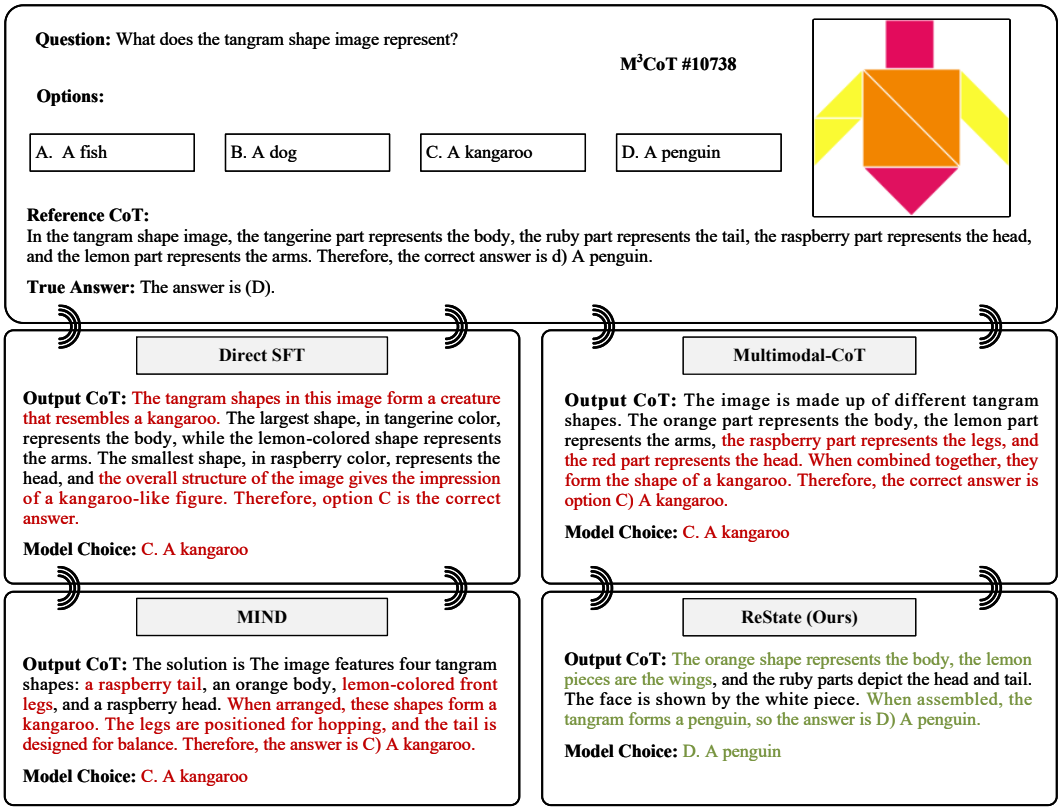


Figure 13: Qualitative comparison on an example of tangram recognition in M^{3}CoT.

connections, corrects the intermediate relation, and reaches the correct answer. In Fig. 16, the initial state correctly computes the densities of the two objects and concludes that neither should sink, but nevertheless produces an inconsistent final choice of "B. 1." Through recursive state feedback, ReState preserves the valid density calculations and flotation judgments while correcting the final answer to "C. 0," which is then maintained in subsequent iterations.

These examples illustrate a typical reasoning-state evolution pattern: *erroneous state* $\rightarrow$ *evidence re-evaluation* $\rightarrow$ *state correction* $\rightarrow$ *correct-state preservation*. Importantly, this process is not merely repeated answer generation. Instead, the model explicitly conditions each update on its current reasoning state, allowing previously generated reasoning to be preserved, reconsidered, or corrected. This qualitative behavior is consistent with the quantitative results in the main text, where most gains are obtained within the first few refinement rounds before performance gradually stabilizes.

The entire refinement process reuses the same MLLM and the same reasoning-state update operator, without additional reflection module, critic, or independent correction model. By incorporating self-generated historical states during training, CRSL exposes the model to states closer to those encountered during recursive inference. At inference, ReState follows a closed loop of "state feedback $\rightarrow$ re-evaluation $\rightarrow$ state update," progressively correcting erroneous states while preserving valid reasoning. These qualitative examples provide intuitive evidence for the iterative self-refinement capability enabled by closed-loop reasoning-state learning.

G More Discussion

G.1 Generalization Evaluation

To further evaluate the cross-dataset transferability of ReState, we directly evaluate models trained on VersaCoT-10K on ScienceQA and M^{3}CoT without additional fine-tuning on the target datasets. We compare ReState with the original pretrained models without SFT and models trained with Direct

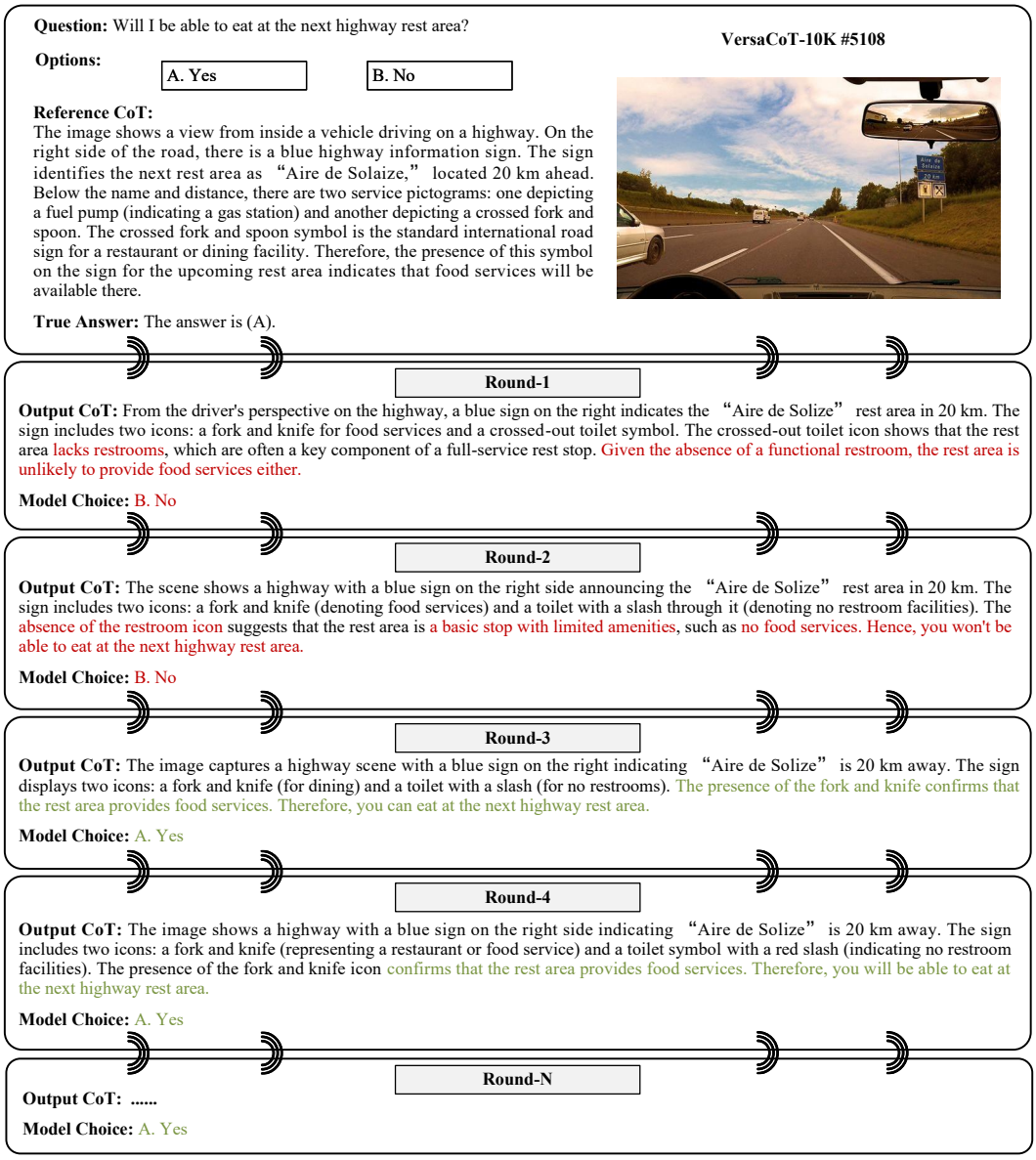


Figure 14: Iterative reasoning-state refinement of ReState on VersaCoT-10K.

SFT on the same VersaCoT-10K data. This setting examines whether the reasoning-state transition capability learned by ReState can transfer to unseen benchmarks beyond the dataset used for training.

As shown in Table 11, under the VersaCoT-10K $\rightarrow$ ScienceQA setting, ReState achieves 85.05% with Qwen2.5-VL-7B and 91.94% with Qwen3.5-9B. These results improve over the corresponding original models by 0.66% and 0.69%, and over Direct SFT by 0.35% and 0.40%, respectively. With Qwen3-VL-8B, ReState reaches 90.10%, which is 0.52% lower than the original model but still higher than the 89.55% obtained by Direct SFT. Overall, on ScienceQA, where the pretrained backbones already exhibit strong performance, ReState largely preserves their original capabilities while providing more consistent cross-dataset transfer than Direct SFT.

The advantage is more evident under the VersaCoT-10K $\rightarrow$ M^{3}CoT setting. ReState achieves 61.17%, 62.12%, and 65.10% with Qwen2.5-VL-7B, Qwen3-VL-8B, and Qwen3.5-9B, respectively, improving over the corresponding original models by 1.38%, 0.69%, and 2.29%, and over Direct SFT by 2.24%, 4.14%, and 1.99%, respectively. Notably, Direct SFT decreases performance relative to the original models by 0.86% on Qwen2.5-VL-7B and 3.45% on Qwen3-VL-8B, whereas ReState

Question: In this food web, which organism contains matter that eventually moves to the earthworm?

ScienceQA #10709

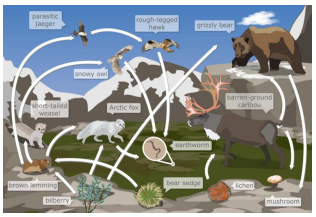
Options:

A. barren-ground caribou

B. mushroom

C. grizzly bear

D. bear sedge



Reference CoT:
Use the arrows to follow how matter moves through this food web. For each answer choice, try to find a path of arrows to the earthworm. The only arrow pointing from the grizzly bear leads to the mushroom. No arrows point from the mushroom to any other organisms. So, in this food web, matter does not move from the grizzly bear to the earthworm. There are three paths matter can take from the bear sedge to the earthworm: bear sedge → brown lemming → Arctic fox → earthworm; bear sedge → brown lemming → snowy owl → earthworm; bear sedge → brown lemming → parasitic jaeger → rough-legged hawk → earthworm. Mushroom. No arrows point from the mushroom to any other organisms. So, in this food web, matter does not move from the mushroom to the earthworm. Barren-ground caribou. There are two arrows pointing from the barren-ground caribou to other organisms. One arrow points to the grizzly bear. The only arrow pointing from the grizzly bear leads to the mushroom. The other arrow pointing from the barren-ground caribou leads to the mushroom. No arrows point from the mushroom to any other organisms. So, in this food web, matter does not move from the barren-ground caribou to the earthworm. There are four paths matter can take from the bilberry to the earthworm: bilberry → Arctic fox → earthworm; bilberry → brown lemming → Arctic fox → earthworm; bilberry → brown lemming → snowy owl → earthworm; bilberry → brown lemming → parasitic jaeger → rough-legged hawk → earthworm.

True Answer: The answer is (D).

Round-1

Output CoT: Use arrow directions to map matter flow. Barren-ground caribou (A) connects to earthworms via two routes: caribou → earthworm and caribou → grizzly bear → earthworm. Mushroom (B) has no outgoing arrows, so matter stops there. Grizzly bear (C) transfers matter to mushroom, which doesn't continue. Bear sedge (D) feeds lemmings, which then transfer to earthworms.

Model Choice: A. barren-ground caribou

Round-2

Output CoT: Track matter flow by following arrows. For each option, find a path to the earthworm. The barren-ground caribou's arrows go to the grizzly bear and mushroom. The grizzly bear's arrow ends at the mushroom, so matter doesn't reach the earthworm. The mushroom has no outgoing arrows, so it doesn't transfer matter. The grizzly bear's arrow ends at the mushroom, so matter doesn't reach the earthworm. The bear sedge's arrow leads to the brown lemming, which connects to the earthworm.

Model Choice: D. bear sedge

Round-3

Output CoT: Follow the arrows to map matter movement. For each option, identify a path to the earthworm. The barren-ground caribou's arrows go to the grizzly bear and mushroom. The grizzly bear's arrow ends at the mushroom, so matter doesn't reach the earthworm. The mushroom has no outgoing arrows, so it doesn't transfer matter. The grizzly bear's arrow ends at the mushroom, so matter doesn't reach the earthworm. The bear sedge's arrow leads to the brown lemming, which connects to the earthworm.

Model Choice: D. bear sedge

Round-N

Output CoT: .....

Model Choice: D. bear sedge

Figure 15: Iterative reasoning-state refinement of ReState on ScienceQA.

improves all three backbones. These results indicate that the benefits of ReState are not confined to the VersaCoT-10K training distribution and can transfer to a different multimodal reasoning benchmark.

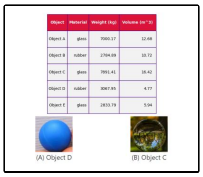
Overall, the cross-dataset results demonstrate promising transferability of ReState. Unlike Direct SFT, which primarily learns a direct mapping from the training data to target outputs, ReState jointly learns reasoning construction, correct reasoning preservation, and erroneous reasoning correction through a shared reasoning-state transition mechanism, while CRSL further exposes the model to self-generated historical states during training. This learning paradigm may encourage reasoning behaviors that are less tied to a specific dataset format and more reusable across tasks. Although ReState does not improve over the original model in every transfer setting, it more consistently preserves or enhances performance across datasets and backbones, with particularly clear gains on M^3CoT. These results further support the cross-dataset applicability of state-centric reasoning and its potential to improve reasoning transfer across heterogeneous multimodal tasks.

Question: In the provided image, how many items from the given choices will submerge underwater?

Options:

A. 2 | B. 1 | C. 0

M^{3}CoT #10416



Reference CoT:
First we should extract the data corresponding to the object mentioned in the image. Object D is made of rubber, weighs 3067.95 kg, and has a volume of 4.77 m³. Object C is made of glass, weighs 7891.41 kg, and has a volume of 16.42 m³. Step 2: At 4 °C, water has a density of approximately 1000 kg/m³. Objects with a lower density than water float, while those with a higher density sink. Step 3: The density of Object D is 3067.95 kg ÷ 4.77 m³ = 643.18 kg/m³, so it floats. Step 4: The density of Object C is 7891.41 kg ÷ 16.42 m³ = 480.60 kg/m³, so it also floats. Therefore, 0 objects sink, and the answer is (C) 0.

True Answer: The answer is (C).

Round-1

Output CoT: Step 1: The image lists Object D (rubber, 3067.95 kg, 4.77 m³) and Object C (glass, 7891.41 kg, 16.42 m³). Step 2: Water has a density of approximately 1000 kg/m³ at 4 °C; objects with lower density float, while objects with higher density sink. Step 3: Object D's density is 3067.95 ÷ 4.77 ≈ 643.18 kg/m³, so it floats. Step 4: Object C's density is 7891.41 ÷ 16.42 ≈ 480.60 kg/m³, so it also floats. Steps 5 – 7: Neither object sinks; 0 items submerge.

Model Choice: B. 1

Round-2

Output CoT: Step 1: The image lists Object D (rubber, 3067.95 kg, 4.77 m³) and Object C (glass, 7891.41 kg, 16.42 m³). Step 2: Water has a density of approximately 1000 kg/m³ at 4 °C; objects with lower density float, while those with higher density sink. Steps 3 – 4: Object D's density is 3067.95 ÷ 4.77 ≈ 643.18 kg/m³, and Object C's density is 7891.41 ÷ 16.42 ≈ 480.60 kg/m³. Both float. Steps 5 – 7: Neither sinks, so 0 objects are submerged. The answer is (C) 0.

Model Choice: C. 0

Round3

Output CoT: Step 1: The image lists Object D (rubber, 3067.95 kg, 4.77 m³) and Object C (glass, 7891.41 kg, 16.42 m³). Step 2: At 4 °C, water has a density of 1000 kg/m³; objects with lower density float, while those with higher density sink. Steps 3 – 4: Object D's density is 3067.95 ÷ 4.77 ≈ 643.18 kg/m³, and Object C's density is 7891.41 ÷ 16.42 ≈ 480.60 kg/m³. Steps 5 – 7: Both objects float, so 0 items sink. The answer is (C) 0.

Model Choice: C. 0

Round-N

Output CoT: .....

Model Choice: C. 0

Figure 16: Iterative reasoning-state refinement of ReState on M^{3}CoT.

G.2 Training–Inference Reasoning-State Gap

As shown in Fig. 1 and Tables 1 to 3 of the main text, the two-stage Multimodal-CoT baseline does not consistently outperform single-stage Direct SFT under matched backbone and training settings. Instead, it performs worse than Direct SFT in eight of the nine dataset-backbone combinations. This degradation is particularly evident on M^{3}CoT. With Qwen2.5-VL-7B and Qwen3-VL-8B, Multimodal-CoT decreases accuracy by 3.10% (from 79.42% to 76.32%) and 2.67% (from 81.62% to 78.95%), respectively. With Qwen3.5-9B, the decrease further reaches 8.80% (from 84.73% to 75.93%). Similar performance drops are observed across all three backbones on ScienceQA and on two of the three backbones on VersaCoT-10K.

This empirical pattern is consistent with the training-inference reasoning-state gap discussed for stage-centric MCoT methods. A typical stage-centric pipeline decomposes reasoning into CoT generation and answer inference. During training, the answer-inference stage is commonly optimized with human-annotated or pre-constructed reasoning chains, whereas at inference time it receives reasoning generated by the preceding model. Such model-generated states may contain omissions, local errors, or reasoning patterns that differ from those observed during training, resulting in a mismatch between the reasoning-state distributions encountered during training and inference. Moreover, existing stage-centric methods (Zhang et al., 2024; Tan et al., 2024a; He et al., 2024; Zheng et al., 2023; Wang et al., 2024a) generally treat intermediate CoTs as stage-specific outputs rather than reusable states

Table 11: Cross-dataset generalization of ReState from VersaCoT-10K to ScienceQA and M^{3}CoT.

Setting	VersaCoT-10K → ScienceQA			VersaCoT-10K → M^{3}CoT		
	Qwen2.5-VL-7B	Qwen3-VL-8B	Qwen3.5-9B	Qwen2.5-VL-7B	Qwen3-VL-8B	Qwen3.5-9B
Origin (No SFT)	84.39 (+0.66)	**90.62** (-0.52)	91.25 (+0.69)	59.79 (+1.38)	61.43 (+0.69)	62.81 (+2.29)
Origin (Direct SFT)	84.70 (+0.35)	89.55 (+0.55)	91.54 (+0.40)	58.93 (+2.24)	57.98 (+4.14)	63.11 (+1.99)
ReState (Ours)	**85.05**	90.10	**91.94**	**61.17**	**62.12**	**65.10**

that can be continuously re-evaluated and refined. Consequently, errors introduced by an upstream reasoning stage may be directly propagated to downstream answer prediction.

In contrast, Direct SFT does not condition a separately optimized downstream stage on an intermediate reasoning state, thus avoiding this source of cross-stage mismatch. The comparison suggests that explicit CoT supervision alone does not necessarily guarantee improved reasoning performance. Effective reasoning also depends on how intermediate reasoning states are modeled and how well their training-time and inference-time distributions are aligned. Although these comparisons do not isolate the reasoning-state gap as the sole cause of the observed degradation, they provide empirical motivation for explicitly modeling model-induced reasoning states during learning.

ReState addresses this issue by reformulating intermediate reasoning from stage-specific outputs into continuously updatable states. URSM unifies reasoning construction, correct reasoning preservation, and erroneous reasoning correction as transitions within a shared reasoning-state space. Building on this unified state-update mechanism, CRSL further incorporates self-generated historical states during training, exposing the model to model-induced states that more closely resemble those encountered during recursive inference. At inference, the same state-update operator recursively feeds back and refines the current reasoning state. Consistent with this design, ReState outperforms Multimodal-CoT across all three datasets and all three backbones. On M^{3}CoT in particular, the gains range from 8.67% (from 76.32% to 84.99%) to 13.07% (from 75.93% to 89.00%).

Overall, these results suggest that effective learning-based MCoT requires both explicit reasoning supervision and consistent reasoning-state modeling across training and inference. By treating intermediate CoTs as iteratively updatable states and incorporating model-induced historical states into closed-loop learning, ReState provides a unified state-centric mechanism for mitigating the training-inference reasoning-state gap and enabling iterative refinement.

G.3 Quantitative Analysis of Generated CoTs

To further evaluate the quality of the generated reasoning processes beyond final-answer accuracy, we employ Qwen3.8-27B (Qwen Team, 2026d) as a unified judge model to assess CoTs produced by different methods on VersaCoT-10K. We evaluate five dimensions: *Overall*, *Correctness*, *Relevance*, *Coherence*, and *Solving Process*. As shown in Tables 12 to 14, ReState achieves the highest *Overall* score across Qwen2.5-VL-7B, Qwen3-VL-8B, and Qwen3.5-9B, while also obtaining the best result on every fine-grained metric. These results indicate that the benefits of ReState extend beyond final-answer accuracy to the quality and consistency of the generated reasoning process.

With Qwen2.5-VL-7B, ReState achieves an *Overall* score of 8.12, improving by 0.12 (from 8.00 to 8.12) over MIND, the strongest competing method. Its *Correctness*, *Relevance*, *Coherence*, and *Solving Process* scores reach 6.82, 9.05, 8.96, and 7.65, respectively. The improvements are more pronounced with Qwen3-VL-8B, where ReState obtains an *Overall* score of 8.85, exceeding the best competing result by 0.36 (from 8.49 to 8.85). Its four fine-grained scores further reach 7.73, 9.47, 9.43, and 8.78, respectively. In particular, the *Solving Process* score improves by 0.49 (from 8.29 to 8.78) over the strongest competing result, suggesting a clearer and more reliable problem-solving trajectory. With Qwen3.5-9B, ReState likewise achieves the highest *Overall* score of 8.51 and consistently outperforms the compared methods across all four fine-grained dimensions. Overall, ReState achieves the strongest result for every evaluated reasoning-quality metric across all three backbones, demonstrating consistent improvements in generated CoT quality.

These improvements are consistent with the design objectives of ReState. URSM jointly models reasoning construction from empty states, correct reasoning preservation and refinement, and erroneous reasoning correction within a shared reasoning-state space. CRSL further exposes the model to self-generated historical states during training, enabling the state-update operator to better handle

Table 12: CoT quality evaluation on VersaCoT-10K with Qwen2.5-VL-7B.

Model	Overall	Correctness	Relevance	Coherence	Solving Process
Origin (No SFT)	7.30 (+0.82)	5.85 (+0.97)	8.44 (+0.61)	8.62 (+0.34)	6.29 (+1.36)
Origin (Direct SFT)	7.82 (+0.30)	6.44 (+0.38)	8.84 (+0.21)	8.61 (+0.35)	7.38 (+0.27)
Multimodal-CoT	7.56 (+0.56)	6.29 (+0.53)	8.37 (+0.68)	8.40 (+0.56)	7.20 (+0.45)
MIND	8.00 (+0.12)	6.63 (+0.19)	9.01 (+0.04)	8.88 (+0.08)	7.49 (+0.16)
ReState (Ours)	**8.12**	**6.82**	**9.05**	**8.96**	**7.65**

Table 13: CoT quality evaluation on VersaCoT-10K with Qwen3-VL-8B.

Model	Overall	Correctness	Relevance	Coherence	Solving Process
Origin (No SFT)	8.21 (+0.64)	6.93 (+0.80)	9.19 (+0.28)	9.15 (+0.28)	7.55 (+1.23)
Origin (Direct SFT)	8.49 (+0.36)	7.25 (+0.48)	9.28 (+0.19)	9.13 (+0.30)	8.29 (+0.49)
Multimodal-CoT	8.27 (+0.58)	7.28 (+0.45)	8.87 (+0.60)	8.85 (+0.58)	8.06 (+0.72)
MIND	8.35 (+0.50)	7.57 (+0.16)	9.13 (+0.34)	8.65 (+0.78)	8.04 (+0.74)
ReState (Ours)	**8.85**	**7.73**	**9.47**	**9.43**	**8.78**

Table 14: CoT quality evaluation on VersaCoT-10K with Qwen3.5-9B.

Model	Overall	Correctness	Relevance	Coherence	Solving Process
Origin (No SFT)	7.97 (+0.54)	6.68 (+0.64)	9.01 (+0.31)	8.78 (+0.51)	7.42 (+0.71)
Origin (Direct SFT)	8.25 (+0.26)	6.81 (+0.51)	9.12 (+0.20)	9.06 (+0.23)	8.00 (+0.13)
Multimodal-CoT	8.17 (+0.34)	7.04 (+0.28)	8.81 (+0.51)	8.91 (+0.38)	7.93 (+0.20)
MIND	8.42 (+0.09)	7.22 (+0.10)	9.25 (+0.07)	9.25 (+0.04)	7.98 (+0.15)
ReState (Ours)	**8.51**	**7.32**	**9.32**	**9.29**	**8.13**

and refine model-induced reasoning states. Consequently, the gains of ReState are reflected not only in final prediction accuracy, but also in stronger reasoning correctness, relevance, coherence, and problem-solving quality. Together with the qualitative examples, these judge-based evaluations provide complementary evidence that state-centric reasoning improves both reasoning outcomes and the intermediate reasoning processes that lead to them.

G.4 Limitations and Future Work

Although ReState consistently improves performance across multiple backbones and multimodal reasoning benchmarks, several limitations remain and warrant further investigation. First, the current experiments primarily focus on multiple-choice reasoning. Although the cross-dataset results demonstrate some transfer capability, applicability to open-ended VQA, video reasoning, and longer-horizon multimodal reasoning remains to be established. Future work could extend ReState to more diverse task formats and model families to evaluate the generality and scalability of state-centric reasoning. Second, ReState achieves iterative self-refinement by repeatedly invoking the same MLLM, maintaining a unified reasoning architecture without additional reflection modules, critics, or independent correction models. However, multiple recursive updates still incur additional inference cost. The current results show that most gains occur in the first few rounds, after which performance gradually stabilizes, indicating that much of the benefit can be obtained with a relatively small number of refinement rounds. Moreover, combining the results in Table 1 and Table 4, even without iterative self-refinement, i.e., under an inference cost comparable to MIND, ReState still improves accuracy by 0.15%-0.90% across the three backbones. This indicates that the gains of ReState do not solely depend on iterative self-refinement. Future work could explore more efficient recurrent reasoning mechanisms, such as Looped Transformers (Giannou et al., 2023), to reduce computational cost while retaining iterative refinement capability. Overall, ReState provides a unified modeling perspective that shifts multimodal reasoning from stage-centric pipelines to state-centric iterative reasoning. Future work could further improve generality, efficiency, and long-horizon refinement stability, extending state-centric reasoning to more open-ended and complex multimodal tasks.