VisMind: Stateful Synthetic Dialogue for Visual Mid-Training

Anonymous ACL submission

Abstract

The mid-training phase has emerged as a critical bridge for enhancing the reasoning capabilities of Vision-Language Models, yet its effectiveness is constrained by the scarcity of high-quality visual reasoning data. While data synthesis offers a promising alternative, conventional stateless synthesis paradigms produce shallow trajectories that quickly succumb to utility plateaus. To overcome this, we propose VisMind, a novel multi-agent collaborative framework that generates deep-reasoning Stateful Synthetic Dialogue (SSD) from visual inputs. By orchestrating iterative problem-solving, where peer student agents engage in bidirectional critique under the ground-truth-anchored guidance of a teacher, VisMind transforms static, fragmented responses into rich, multi-turn logical derivations. Utilizing this interactive paradigm, we construct a dense mid-training corpus of approximately 1.14B tokens. Extensive experiments across the Qwen2-VL and InternVL3 model families demonstrate that applying this corpus via Continued Pretraining consistently outperforms a representative stateless synthesis approach and elevates standard baselines by up to a 2.24% absolute average gain across comprehensive mathematical and logical benchmarks. Furthermore, in-depth analysis reveals that the SSD paradigm exhibits continuous linear scaling without early saturation, fosters intrinsic reasoning generalization, and establishes a robust cognitive foundation that enhances the efficacy of subsequent Reinforcement Learning with Verifiable Rewards.

1 Introduction

Vision-Language Models (VLMs) have advanced multimodal intelligence, yet robust visual reasoning remains challenging (Team et al., 2026; Bai et al., 2025; Achiam et al., 2023). Conventional pipelines establish foundational knowledge and cross-modal alignment through pre-training on massive data (Liu et al., 2023), transitioning to post-training via Reinforcement Learning with Verifiable Rewards (RLVR) (Guo et al., 2025) and Supervised Fine-Tuning (SFT) (Ouyang et al., 2022) to strengthen reasoning and human alignment (Zhang et al., 2024a; Chu et al., 2025). Bridging these stages, mid-training alters data distributions to target objectives including domain expansion and data quality (Wang et al., 2025b; Tu et al., 2025). As a primary approach, Continued Pretraining (CPT) (Gururangan et al., 2020; Mo et al., 2025) applies next-token prediction on refined, domain-specific corpora to internalize specialized knowledge. Integrating high-quality instruction data during the CPT annealing phase enhances base models by aligning distributions and mitigating noise to prime them for effective post-training (Parmar et al., 2024; Hu et al., 2024). However, this synergy is impeded by the dearth of high-quality visual reasoning data, which remains difficult to curate due to complex spatial-logical alignments (Wang et al., 2024b; Kumichev et al., 2024).

To mitigate data scarcity, research increasingly adopts data synthesis (Akter et al., 2024; Yang et al., 2024; Ding et al., 2023). While promising, mainstream synthetic methods (e.g., response diversification, query expansion, and retrospective enhancement (Chen et al., 2025b)) rely on stateless inference, where a solitary completion pass from an initial prompt yields basic Query-Answer (QA) pairs or shallow *Short Chain-of-Thought* (CoT) (Wei et al., 2022) trajectories (see Figure 1, Top). Restricted by limited seeds, these stateless paradigms struggle to transcend the model's primary probability space even with high-temperature sampling, failing to exploit the generator's reasoning potential and causing data utility plateaus and stagnant scaling (Madaan et al., 2023). Trained on such data, models learn trivial heuristics rather than internalizing the complex, multi-hop logic and reflective error-correction required for robust visual reasoning (Chen et al., 2025a). Scaling volume alone

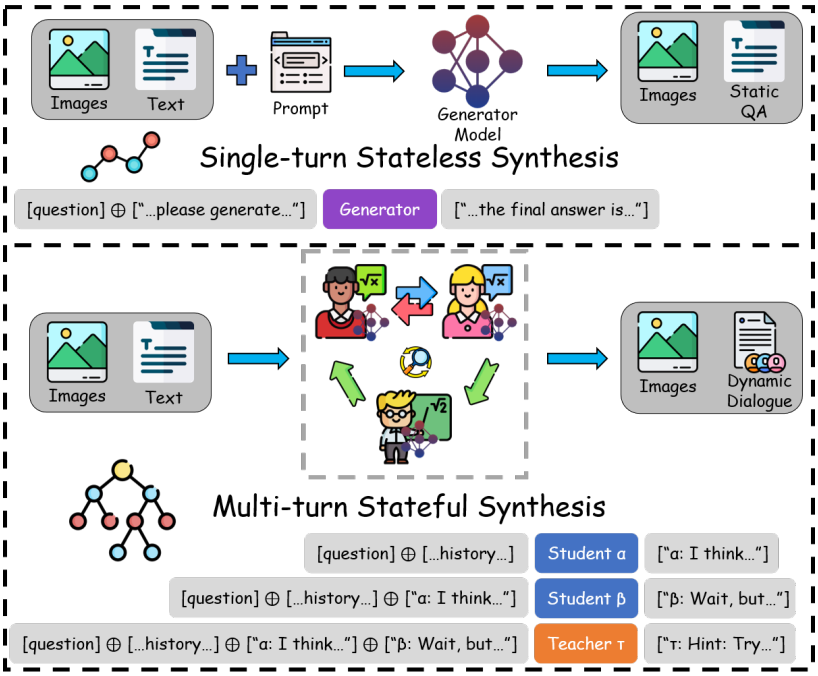


Figure 1: Data synthesis paradigms. **Top:** Stateless synthesis employs single-turn inference for static QA pairs with shallow reasoning topologies. **Bottom:** VisMind orchestrates iterative multi-agent interactions for dynamic dialogues with deep logical trajectories.

cannot resolve these deficiencies, which necessitates a transition toward stateful generation where dynamic interaction cycles push the generator to construct deeper logical trajectories.

To overcome these limitations, we propose VisMind,[1] a novel multi-agent framework generating sophisticated Stateful Synthetic Dialogue (SSD) (see Figure 1, Bottom). Moving beyond the solitary completion passes of stateless generation, VisMind simulates iterative problem-solving on visual seeds, with peer students collaboratively exploring multimodal queries. Instead of static outputs, they engage in bidirectional critique by proposing, challenging, and evaluating reasoning steps. For validation, a ground-truth-aware teacher provides corrective hints rather than direct solutions, steering them to refine logic across multiple cycles. This supervised collaboration replaces fragmented *Short CoT* answers with multi-turn *Long CoT* trajectories embedding alternative path exploration, reflective behaviors, and joint error correction. By fully exploiting the generator's potential via stateful interaction, the SSD paradigm yields a high-quality synthetic corpus infused with rich knowledge and reasoning behaviors to strengthen base models during mid-training.

[1]The code is available at https://anonymous.4open.science/r/VisMind-E816.

Through VisMind, we synthesize a 1.14B-token SSD corpus. Four VLM variants mid-trained on this corpus via CPT, followed by standard SFT, consistently outperform their pure SFT baselines, achieving up to 2.24% absolute average gains across comprehensive reasoning benchmarks. Crucially, under identical configurations, this stateful approach surpasses stateless QA synthesis by up to 2.23%, validating the superiority of interactive dynamics over static generation. Furthermore, SSD avoids single-pass homogenization, exhibiting high semantic diversity and continuous data utility scaling. Probing intrinsic reasoning capabilities verifies that this paradigm fosters genuine cognitive generalization rather than superficial memorization. Ultimately, internalizing these complex derivations establishes a robust cognitive foundation, preventing strategy collapse during subsequent RLVR.

In summary, our main contributions include:

- **Stateful synthesis paradigm.** We introduce VisMind, an open-source multi-agent framework that fundamentally shifts visual data synthesis from static, single-turn inference to dynamic, self-refining dialogue. This pipeline provides a scalable mechanism to generate dense, long-horizon reasoning trajectories, effectively utilized to synthesize our 1.14B-token mid-training corpus.

- **Empirical scaling dynamics.** We establish that interactive stateful dynamics successfully break the data utility plateaus inherent to conventional stateless synthesis. By validating continuous data utility scaling and bypassing response homogenization, we confirm that teacher-guided peer critique offers a sustainable pathway for scaling reasoning data.

- **Foundation for alignment pipeline.** We reveal a critical synergy between mid-training and reinforcement learning. By demonstrating that internalizing complex derivations fosters intrinsic cognitive generalization, we identify that a robust mid-trained reasoning foundation is a vital prerequisite for preventing policy degradation and sustaining effective exploration during RLVR.

2 The VisMind framework

VisMind orchestrates SSD through a three-stage workflow where multi-agent interaction transforms static visual seeds into multi-turn trajectories, followed by rule-based filtering, leading to training integrating SSD via CPT before SFT on the seeds.

2.1 Stateful synthetic dialogue

We formulate SSD generation as a supervised, multi-turn Markovian interaction among three specialized agents: two peer-reasoning students (α and β) engaging in bidirectional critique, and a ground-truth-aware teacher (τ).

2.1.1 Problem formulation

Given a visual reasoning query $Q = (x, v)$, where x represents the textual instruction and v represents the associated visual modalities positioned as a fixed leading context, along with a corresponding ground-truth solution GT, our objective is to synthesize a valid dialogue trace $\mathcal{D} = \langle Q, \mathcal{R} \rangle$. Here, $\mathcal{R} = (\rho_1, \rho_2, \ldots, \rho_K)$ represents the ordered sequence of K utterances generated by the agents. We define a macroscopic exchange cycle t between the two student agents, incrementing t solely upon the completion of β's utterance.

We define the active agent state space as $\mathcal{S} = \{\alpha, \beta, \tau\}$, supplemented by a terminal state ξ. The generation of $\mathcal{R}$ is a conditional sequential decision process, initialized at $\sigma_1 = \alpha$. To emulate independent problem-solving, GT is strictly isolated from the student agents. The probability distribution of the utterance ρ_k and the next-speaker transition σ_{k+1} at step k is conditioned on the current agent σ_k and the dialogue histories $\mathcal{R}_{<k} = (\rho_1, \ldots, \rho_{k-1})$ and $\mathcal{R}_{\leq k} = (\rho_1, \ldots, \rho_k)$:

$$
P(\rho_k, \sigma_{k+1} \mid Q, \mathcal{R}_{<k}, \sigma_k, GT) =
\begin{cases}
P_\alpha(\rho_k \mid Q, \mathcal{R}_{<k}) \cdot \mathbb{I}(\sigma_{k+1} = \beta), \\
\qquad \text{if } \sigma_k = \alpha \\
P_\beta(\rho_k \mid Q, \mathcal{R}_{<k}) \cdot \Pi_\beta(\sigma_{k+1} \mid Q, \mathcal{R}_{\leq k}), \\
\qquad \text{if } \sigma_k = \beta,\ \sigma_{k+1} \in \{\alpha, \tau\} \\
P_\tau(\rho_k \mid Q, \mathcal{R}_{<k}, GT) \cdot \Pi_\tau(\sigma_{k+1} \mid Q, \mathcal{R}_{\leq k}, GT), \\
\qquad \text{if } \sigma_k = \tau,\ \sigma_{k+1} \in \{\alpha, \xi\}
\end{cases}
\tag{1}
$$

where $\mathbb{I}(\cdot)$ denotes an indicator function, Π_β and Π_τ denote routing policies over admissible next states, and $P_\alpha, P_\beta, P_\tau$ denote the generation distributions parameterized by a single underlying generator model, from which distinct reasoning and routing behaviors are elicited by conditioning on role-specific system prompts (see Appendix A). The framework samples trajectories that are accepted by the teacher, retaining the valid trace $\mathcal{D}^* = \langle Q, \mathcal{R} \rangle$ if and only if the final transition reaches the terminal state ξ within the maximum exchange cycle limit $T_{\max}$:

$$
\begin{aligned}
&\mathcal{R} \sim P(\mathcal{R} \mid Q, GT), \\
&\mathcal{D}^* = \langle Q, \mathcal{R} \rangle \quad \text{s.t.} \quad \sigma_{K+1} = \xi \wedge t \leq T_{\max}
\end{aligned}
\tag{2}
$$

2.1.2 Multi-agent collaborative dynamics

Collaborative exploration ($\alpha \leftrightarrow \beta$). Students α and β iteratively generate reasoning steps (ρ_α, ρ_β) while exploring alternative reasoning paths. This peer-to-peer loop is driven by bidirectional critique, where each student evaluates the counterpart's previous derivation by explicitly supporting or challenging its logic. This lateral exploration mitigates the exploration deficit inherent to binary student-teacher setups, preventing premature convergence on suboptimal branches prior to supervised intervention. While both agents act as equal peers in reasoning and verification, the framework imposes an asymmetric routing mechanism to manage dialogue progression, where β uniquely acts as a conditional router. Based on its internal verification of the augmented context $(Q, \mathcal{R}_{\leq k})$, β determines the state transition $\sigma_{k+1} \in \{\alpha, \tau\}$ by returning control to α to rectify identified mistakes, or invoking the teacher when the derivation is deemed fully correct and requires final validation.

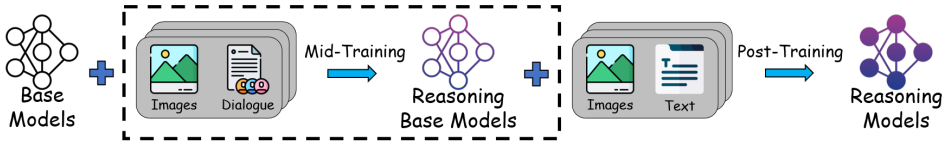


Figure 2: The VisMind training pipeline. Base models are mid-trained on the synthesized SSD corpus to fortify foundational reasoning, followed by post-training on the original visual instructions for final alignment.

Supervised intervention (τ). When invoked ($\sigma_k = \tau$), the teacher evaluates the accumulated history $\mathcal{R}_{<k}$ for reasoning accuracy using its privileged access to GT. Upon detecting errors, misconceptions, or incomplete logic, τ generates constructive feedback and hints (ρ_τ) rather than direct answers. By routing the conversation back to the students ($\sigma_{k+1} = \alpha$), this intervention compels them to integrate the feedback to refine their derivation, effectively preventing cascading errors without sacrificing cognitive depth.

Termination and convergence (ξ). The interaction cycle continues until τ confirms the derivation is congruent with GT, issuing the termination signal ($\sigma_{k+1} = \xi$). If t reaches $T_{\max}$ without validation, the trace is discarded. Repeated execution across diverse queries yields the raw synthetic corpus $\mathcal{C}_{raw} = \{\mathcal{D}_1^*, \ldots, \mathcal{D}_N^*\}$, comprising N successfully verified trajectories. Collectively, the stateful interaction paradigm overcomes the limitations of solitary inference, yielding a corpus of complex, multi-turn reasoning trajectories ready for filtering.

2.2 Data filtering

To ensure the high fidelity of the raw synthetic corpus $\mathcal{C}_{raw}$, we implement a three-stage rule-based filtering pipeline. The first two stages act as rejection filters, discarding the entire dialogue trace $\mathcal{D}^*$ upon failure, while the final stage performs content-level sanitization.

Language verification. Although our multi-agent prompts constrain the models to communicate exclusively in English, we utilize Unicode block matching to reject any dialogue $\mathcal{D}^*$ containing non-English character sets, excluding mathematical symbols. Out-of-domain scripts indicate a failure in instruction following, rendering the interaction unreliable.

Repetition detection. Extended multi-turn generation makes VLMs susceptible to inference degradation and pathological text loops. We apply a sequential pattern detector to discard trajectories exhibiting consecutive triple repetitions of any word or phrase, thereby preserving the lexical diversity and logical progression of the corpus.

Format cleaning. In validated dialogues, all internal markers managing Markovian state transitions and termination are systematically stripped.[2] The remaining outputs are concatenated into clean, contiguous conversational trajectories suitable for the subsequent mid-training phase.

2.3 Training pipeline

As illustrated in Figure 2, VisMind modifies the conventional VLM training trajectory by inserting a strategic mid-training phase between initial pre-training and post-training, executed through a three-step pipeline.

Seed data preparation. We utilize standard visual instruction data, typically reserved for the SFT stage, as foundational seeds. Each raw instance, comprising a multimodal query Q and a static answer A, is processed through the SSD framework. By sampling each seed n times and applying the rejection filters, we transcend the static A to extract up to n diverse, multi-turn reasoning trajectories $\mathcal{R}$, providing the structural complexity required for mid-training.

Mid-training via CPT. The pretrained base model is mid-trained on the filtered SSD corpus using the CPT objective. Processing these long-horizon conversational trajectories enables the model to internalize the underlying cognitive logic and multi-turn reasoning patterns embedded within the SSD trajectories, thereby yielding a robust reasoning base model.

Post-training via SFT. Finally, the reasoning base model undergoes standard SFT using the original visual instruction seeds. With foundational logic solidified during CPT, this stage focuses on aligning the model's reasoning behaviors with diverse human instructions and specific formatting constraints to produce the final reasoning model.

[2] #TO_STUDENT_ALPHA#, #TO_TEACHER#, and #END_CONVERSATION#.

3 Experiments

3.1 Setup

Data synthesis. We utilize 118K multimodal instructions from the `visualwebinstruct` subset of the VisualWebInstruct (Jia et al., 2025) dataset as foundational seeds. To generate the SSD corpus, the aligned InternVL3-14B model is deployed as the agent generator via the vLLM (Kwon et al., 2023) framework. To govern the multi-agent interaction, we set $T_{\max}$ to 5 and sample each seed instance 8 times with a temperature of 0.7, a top-p of 0.8, and a maximum of 8192 new tokens per generation.

Training configuration. Experiments are conducted on $4\times$ NVIDIA A800 80GB GPUs within a single node using the SWIFT (Zhao et al., 2025) framework. We apply the CPT and subsequent SFT pipeline to the pretrained base models of the Qwen2-VL (2B, 7B) (Wang et al., 2024c) and InternVL3 (2B, 8B) (Zhu et al., 2025) series.[3] Training is optimized using DeepSpeed (Rasley et al., 2020) ZeRO-3 with a learning rate of 1×10^{-5}, a warmup ratio of 0.05, and 3 epochs. The global batch size is 8 for 2B models and 4 for 7B/8B models, with a gradient accumulation step of 2. For sequence processing, we enable padding-free packing and FlashAttention (Dao et al., 2022), restricting the maximum token length per sample to 4096. These configurations are applied uniformly across both training stages. For Qwen-series models, we implement specific constraints by restricting the maximum pixel count to 262,144 due to computational limits, equating to approximately 334 visual tokens under the inherent 28×28 pixel-per-token resolution, and, following the official implementation, by freezing the Vision Transformer (ViT) (Dosovitskiy et al., 2020) and the vision-language adapter during the SFT stage. Unspecified hyperparameters follow default framework settings, and all parameters not explicitly mentioned as frozen remain trainable throughout the training process.

Baselines and evaluation. Our primary baselines consist of models trained exclusively via standard SFT on the original visual instruction seeds. To contrast our framework with conventional stateless synthesis, we also evaluate a comparative approach using QA pairs generated via query expansion (Chen et al., 2025b) (see Appendix B), utilizing identical generation configurations and training pipelines as those used in VisMind but differing solely in the synthesized mid-training data. To assess visual, mathematical, and logical reasoning capabilities, we adopt the benchmark selection from InternVL3, covering MathVerse Mini (Zhang et al., 2024b), MathVision (Wang et al., 2024a), and MathVista Mini (Lu et al., 2023) for mathematical reasoning, supplemented by DynaMath (Zou et al., 2024), WeMath (Qiao et al., 2025), and LogicVista (Xiao et al., 2024) for logical reasoning.

[3] Selected due to hardware constraints and the scarcity of open-source *base* VLMs, as most exclusively release post-trained checkpoints.

3.2 Results

Scale and structural expansion. The agentic generation framework transforms static visual instructions into structurally complex reasoning trajectories. The initial `visualwebinstruct` seed subset contains 60.73M tokens, averaging 512.60 tokens per sequence. Given identical seeds and sampling, single-pass stateless inference yields 485.54M tokens, averaging 514.33 tokens per sequence. In contrast, following multi-agent generation and rule-based data filtering, yielding a 77% trajectory retention rate, the SSD corpus scales to 1.14B tokens, tripling the average trajectory length to 1560.32 tokens. This volumetric and structural expansion establishes the dense, long-horizon data foundation required for effective mid-training (see Appendix H for a comparative example).

Multi-agent interaction dynamics. Detailing the macroscopic student-teacher loop, subplots (b) and (c) of Appendix Figure 4 reveal that while 49.6% of dialogues converge in a single cycle, the remainder necessitate sustained problem-solving, with 45.1% requiring multiple teacher interventions to rectify flawed logic and 9.1% reaching the limit of five exchange cycles. Isolating micro-level peer dynamics prior to teacher invocation, subplot (d) demonstrates that although 91.9% of internal exchanges conclude in two utterances, 8.1% extend to four or six, confirming the routing agent β enforces autonomous revisions before external validation. These macro and micro behaviors jointly shape the total utterance count K in subplot (a), where the standard $\alpha \rightarrow \beta \rightarrow \tau$ cadence peaks at multiples of three (e.g., $K \in \{3, 6, 9\}$) and structural variance from autonomous peer rejections systematically generates intermediate dialogue lengths (e.g., $K \in \{5, 8\}$). These interaction metrics demonstrate that the stateful synthesis framework encodes iterative reasoning chains, alternative path exploration, and reflective error correction into the SSD.

Model	Method	MathVerse	MathVision	MathVista	DynaMath	WeMath	LogicVista	Average
Qwen2-VL-2B	Baseline	11.68%	13.95%	39.80%	23.69%	10.95%	**32.66%**	22.12%
	QA	11.78%	14.01%	**43.10%**	25.49%	11.33%	29.08%	22.47%
	VisMind	**14.90%**	**14.90%**	42.60%	**27.64%**	**13.90%**	29.31%	**23.88%**
Qwen2-VL-7B	Baseline	20.43%	20.43%	56.60%	36.65%	22.76%	40.72%	32.93%
	QA	22.51%	21.25%	55.40%	36.63%	22.00%	39.82%	32.94%
	VisMind	**24.87%**	**22.14%**	**59.00%**	**37.37%**	**25.81%**	**41.83%**	**35.17%**
InternVL3-2B	Baseline	19.72%	19.28%	**51.20%**	33.43%	17.90%	35.57%	29.52%
	QA	20.46%	19.11%	49.40%	35.13%	22.00%	32.89%	29.83%
	VisMind	**22.72%**	**20.26%**	48.80%	**36.41%**	**22.38%**	**37.58%**	**31.36%**
InternVL3-8B	Baseline	**25.28%**	25.82%	**63.60%**	45.23%	33.43%	42.73%	39.35%
	QA	24.24%	27.27%	59.60%	45.79%	**36.48%**	42.51%	39.31%
	VisMind	24.06%	**27.93%**	60.20%	**46.81%**	34.95%	**44.52%**	**39.74%**

Table 1: Performance comparison across mathematical and logical reasoning benchmarks. **Baseline** represents base models trained solely via SFT on the visual instruction seeds. **VisMind** denotes base models mid-trained on the SSD corpus, followed by SFT on identical seeds. **QA** signifies base models mid-trained on the QA pairs synthesized via the comparative stateless query expansion approach.

Comparison of reasoning performance. Table 1 compares VisMind against standard SFT baselines and the comparative QA approach. Integrating the SSD corpus via CPT consistently improves average accuracy across all evaluated models, yielding absolute average gains of 1.76%, 2.24%, 1.84%, and 0.39% over the standard baselines for Qwen2-VL-2B, Qwen2-VL-7B, InternVL3-2B, and InternVL3-8B, respectively. These improvements span complex reasoning tasks, with consistent increases on benchmarks such as MathVision and DynaMath. Furthermore, while the stateless QA method occasionally provides marginal improvements over the baseline, VisMind consistently outperforms it in average accuracy across all evaluated models despite sharing identical seeds and training pipelines, achieving margins up to +2.23% and +1.53% on the Qwen2-VL-7B and InternVL3-2B models, respectively. Consequently, these empirical gains validate the efficacy of the interactive generation paradigm, establishing that training on multi-turn trajectories substantially bolsters accuracy across complex visual reasoning tasks.

3.3 Analysis

How do scalability and diversity compare between stateful and stateless paradigms?

To dissect the mechanisms underlying the observed reasoning enhancements, we evaluate data utility by scaling the synthesized samples per seed up to eight during the mid-training of the InternVL3-2B model. As illustrated in Figure 3(a), the model mid-trained on the SSD corpus exhibits continuous, near-linear accuracy improvements as the sampling scale increases, showing no saturation at the maximum evaluated scale. In contrast, the utility of the QA corpus plateaus and degrades beyond four samples. This divergence indicates that while stateless inference rapidly exhausts the utility of single-pass outputs, stateful interaction continuously distills rich knowledge and diverse reasoning behaviors from identical seeds.

To quantify the semantic variance driving this persistent utility, we compute the intra-group cosine similarity, detailed in Appendix C, of the outputs across a shared random subset of 1,000 instruction seeds using the Qwen3-Embedding-0.6B (Zhang et al., 2025) model. Figure 3(b) demonstrates that the SSD distribution shifts toward lower similarity, achieving a reduced average of 0.220 compared to 0.261 for the QA method. Furthermore, the stateful paradigm suppresses highly redundant outputs, lowering the maximum similarity bound from 0.489 to 0.392. This measurable reduction confirms that dynamically accumulating dialogue history continuously alters the conditioning context. This progressively expanding prompt inherently diversifies the generator's search space, bypassing the response homogenization of static inputs to sustain data utility.

Does VisMind foster genuine reasoning or merely memorize synthetic patterns?

To answer this, we evaluate the framework on two independent benchmarks, where MME-CoT (Jiang et al., 2025) assesses thought process quality, robustness, and efficiency, and MME-Reasoning

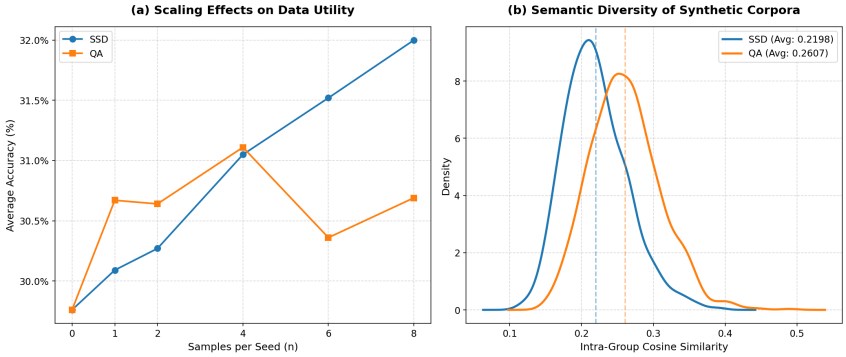


Figure 3: Data utility scaling and semantic diversity of the SSD and QA corpora, represented in blue and orange respectively. **(a)** Scaling effects on data utility: evaluates average reasoning accuracy across synthesized samples per seed n, tracking performance evolution as the sampling scale increases. **(b)** Semantic diversity of synthetic corpora: presents the probability density of intra-group cosine similarities from text embeddings, where lower similarity denotes higher diversity among multiple outputs from identical seeds, and dashed lines mark the group averages.

(Yuan et al., 2025) tests the breadth of logical reasoning across diverse scenarios. As detailed in Appendix Table 3, integrating the QA corpus synthesized via stateless inference typically degrades the structural integrity of the thought process relative to the standard baseline. In contrast, VisMind restores and enhances performance, marked by an increase in the Qwen2-VL-2B model's Overall score from 0.266 for the stateless comparative approach to 0.346. High scores in CoT Relevance and Quality suggest that rather than memorizing static synthetic templates that impede the QA models, the stateful interaction paradigm enables the model to internalize the dynamic reasoning behaviors required to transcend surface-level heuristics.

Appendix Table 4 corroborates this conclusion, showing that while the stateless QA approach diminishes reasoning capabilities across the evaluated model configurations, VisMind demonstrates consistent gains across core logical paradigms including inductive, deductive, and abductive reasoning. For InternVL3-8B, VisMind exhibits substantial advantages over the QA approach in Causal Chain of 43.06 versus 33.33 and Deductive reasoning of 32.03 versus 27.45. The performance gap on these independent benchmarks indicates that the iterative refinement within the SSD trajectories translates into generalized cognitive capabilities. These findings demonstrate that the structural complexity of the stateful paradigm imparts the multi-faceted logic necessary to establish a robust cognitive foundation, effectively bolstering the model's reasoning generalization to outperform both standard direct post-training and conventional stateless synthesis in complex reasoning depth.

Can VisMind provide a more robust foundation for RLVR?

To evaluate the synergy between mid-training paradigms and RLVR, we apply Group Relative Policy Optimization (GRPO) (Guo et al., 2025) to the respective SFT checkpoints of the evaluated Qwen2-VL configurations using approximately 9k reasoning samples from the OpenVLThinker (Deng et al., 2025) dataset. Driven by an external verification reward mechanism (see Appendix E), the optimization spans a maximum of 500 steps, equating to roughly seven epochs, with checkpoints preserved at 50-step intervals. As illustrated in Table 2, GRPO induces a performance divergence. VisMind leverages the reward signals, advancing its 2B and 7B models from their initial SFT averages of 23.88% and 35.17% to 24.54% and 35.81%, respectively. Conversely, the standard baseline and the comparative QA approach exhibit post-RLVR degradation, evidenced by the 7B model trained via the QA approach regressing from 32.94% to 32.38%. This divergence demonstrates the dependency of RLVR on the foundational robustness of the initial policy. When the policy relies on shallow heuristics inherent to standard SFT or data synthesized via stateless inference, the reinforcement learning feedback mechanism amplifies logical inconsistencies. Reward signals consequently reinforce erroneous trajectories, biasing the policy toward unreliable reasoning patterns. These findings confirm that the stateful interaction paradigm establishes the cognitive stability required to sustain RLVR exploration, enabling continuous performance scaling rather than strategy collapse.

Model	Method (+GRPO)	MathVerse	MathVision	MathVista	DynaMath	WeMath	LogicVista	Average
Qwen2-VL-2B	Baseline	9.75%	12.96%	42.00%	23.17%	10.29%	27.96%	21.02%
	QA	10.03%	13.19%	43.50%	25.31%	10.48%	27.52%	21.67%
	VisMind	**15.30%**	**14.67%**	**44.00%**	**27.17%**	**15.90%**	**30.20%**	**24.54%**
Qwen2-VL-7B	Baseline	19.47%	19.64%	58.30%	36.47%	21.62%	40.72%	32.70%
	QA	20.46%	19.74%	59.40%	36.11%	22.76%	35.79%	32.38%
	VisMind	**25.41%**	**22.07%**	**61.20%**	**37.52%**	**25.90%**	**42.73%**	**35.81%**

Table 2: Mathematical and logical reasoning performance of the Qwen2-VL series optimized via RLVR. GRPO is applied uniformly to the SFT checkpoints of the standard baseline, the comparative QA approach, and the VisMind framework to assess their synergy with reinforcement learning and the stability of their initial cognitive foundations.

4 Related work

Visual reasoning and mid-training. While VLMs achieve broad multimodal alignment (Zhang et al., 2024a), translating perception into complex reasoning remains challenging (Wu et al., 2023). Recent paradigms bridge this gap via mid-training (Tu et al., 2025). Executing CPT on curated corpora mitigates data saturation and systematically enhances logical competencies (Mo et al., 2025). Crucial for internalizing specialized knowledge like biomedical diagnostics (Zhang et al., 2026) and embodied priors (Yu et al., 2026), mid-training depends on high-quality reasoning data, necessitating robust synthesis frameworks.

Stateless data synthesis. To mitigate data scarcity and repetition (Hao et al., 2025), foundational synthesis approaches extract entities for domain corpora (Yang et al., 2024) and construct concept graphs to scale instruction tuning (Tang et al., 2024). For complex reasoning, methods bootstrap questions via backward reasoning (Yu et al., 2023) and key-point matrices (Huang et al., 2025). To increase structural complexity, frameworks construct preference trees (Yuan et al., 2024), expand queries (Chen et al., 2025b), and transform text into simulated dialogues (Akter et al., 2024). However, these stateless paradigms rely on solitary inference passes, yielding static trajectories that restrict search spaces and precipitate utility plateaus. VisMind breaks these plateaus through stateful interaction, diversifying exploration via accumulated dialogue history.

Multi-agent collaboration. Multi-agent frameworks scale inference compute by synthesizing general instructions (Tang et al., 2025), automating trajectory verification (Sengupta et al., 2024), and scaling workflows via decentralized architectures (Wang et al., 2025a). For deeper reasoning, studies leverage diverse debates (Hegazy, 2024), tree search coordination (Ye et al., 2024), structured roles (Li et al., 2024), and textual critiques (Wang et al., 2023). Although collaborative agents decompose multimodal queries during inference (Rajput et al., 2025), current paradigms risk unguided reasoning drifts and overlook mid-training. VisMind addresses these gaps via a peer-teacher triad. Coupling bidirectional peer critique for lateral exploration with supervised teacher intervention for factual grounding prevents premature convergence, synthesizing SSD to embed complex derivations.

5 Conclusion

In this paper, we introduced VisMind, a multi-agent collaborative framework designed to overcome the limitations of stateless data synthesis during visual mid-training. By orchestrating bidirectional peer critique and ground-truth-anchored supervised intervention, the proposed stateful paradigm transforms static visual instructions into complex, dynamic reasoning trajectories. This structured interaction mitigates the exploration deficit inherent to solitary inference, yielding a high-quality SSD corpus. Empirical evaluations demonstrate that integrating this corpus via CPT consistently improves diverse VLMs over standard baselines, while outperforming comparative stateless approaches across comprehensive mathematical and logical benchmarks. Furthermore, our analysis reveals that the stateful paradigm bypasses traditional data utility plateaus, exhibits linear scaling, fosters intrinsic reasoning generalization, and establishes the foundational robustness required to sustain subsequent RLVR without strategy collapse. While the three-agent formulation provides an effective approach, future work will explore structural ablations across agent configurations and lightweight interaction mechanisms to reduce the computational overhead of the synthesis process.

Limitations

While VisMind demonstrates substantial empirical success, we acknowledge three primary limitations in our current study.

Computational cost and experimental scale. Multi-agent interaction inherently incurs higher computational costs than single-turn inference, restricting the scale of our empirical validation. Future verification requires scaling the seed data volume, comparing a broader range of synthesis methodologies, and evaluating across more diverse model families and larger parameter scales.

Agent configuration and homogeneity. Our framework relies on a fixed triad of two peer students and one teacher, instantiated by a single generator model. We have not conducted ablation studies on the number of interacting agents to optimize the trade-off between generation efficiency and trajectory quality. Additionally, employing heterogeneous setups where different frontier models assume distinct agent roles remains an unexplored direction for introducing cognitive diversity.

Interpretability of data efficacy. Although empirical results confirm that the SSD corpus significantly enhances reasoning, the fine-grained mechanisms underlying this improvement remain opaque. Our analysis relies on macro-level benchmark scores rather than investigating exactly why these dynamic trajectories yield high-quality training signals. Future work requires mechanistic interpretability to explicitly attribute downstream performance gains to specific multi-agent behaviors.

Ethical considerations

This research does not involve human subjects, crowdsourcing, or the collection of personally identifiable information or offensive content, utilizing only publicly available datasets for data synthesis. The proposed VisMind framework focuses strictly on enhancing the logical and mathematical reasoning capabilities of existing VLMs, and does not inherently introduce new ethical risks, societal biases, or malicious applications beyond those generally associated with foundational multimodal models. Furthermore, in accordance with ACL guidelines regarding the use of generative AI, we transparently disclose that large language models were utilized as assistive tools strictly for manuscript language polishing and experimental code generation, with the authors retaining full responsibility for the final scientific content and integrity.

References

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

Syeda Nahida Akter, Shrimai Prabhumoye, John Kamalu, Sanjeev Satheesh, Eric Nyberg, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Mind: Math informed synthetic dialogues for pretraining llms. *arXiv preprint arXiv:2410.12881*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025a. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Zui Chen, Tianqiao Liu, Mi Tian, Qing Tong, Weiqi Luo, and Zitao Liu. 2025b. Advancing mathematical reasoning in language models: The impact of problem-solving data, data synthesis methods, and training stages. *arXiv preprint arXiv:2501.14002*.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 10818–10838.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. *arXiv preprint arXiv:2503.17352*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words:

Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8342–8360.

Xintong Hao, Ruijie Zhu, Ge Zhang, Ke Shen, and Chenggang Li. 2025. Reformulation for pretraining data augmentation. *arXiv preprint arXiv:2502.04235*.

Mahmood Hegazy. 2024. Diversity of thought elicits stronger reasoning capabilities in multi-agent debate frameworks. *arXiv preprint arXiv:2410.12853*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, and 1 others. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Yiming Huang, Xiao Liu, Yeyun Gong, Zhibin Gou, Yelong Shen, Nan Duan, and Weizhu Chen. 2025. Key-point-driven data synthesis with its enhancement on mathematical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24176–24184.

Yiming Jia, Jiachen Li, Xiang Yue, Bo Li, Ping Nie, Kai Zou, and Wenhu Chen. 2025. Visualwebinstruct: Scaling up multimodal instruction data through web search. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1393.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, Bo Zhang, Chaoyou Fu, Peng Gao, and Hongsheng Li. 2025. MME-CoT: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 27793–27830.

Gleb Kumichev, Pavel Blinov, Yulia Kuzkina, Vasily Goncharov, Galina Zubkova, Nikolai Zenovkin, Aleksei Goncharov, and Andrey Savchenko. 2024. Medsyn: Llm-based synthetic medical text generation framework. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 215–230. Springer.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Siyue Li, Xiaofan Zhou, Zhizhong Wu, Yuiian Long, and Yanxin Shen. 2024. Strategic deductive reasoning in large language models: A dual-agent approach. In *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, pages 834–839. IEEE.

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

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

Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, and Wenhu Chen. 2025. General-reasoner: Advancing llm reasoning across all domains. *arXiv preprint arXiv:2505.14652*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594.

Kaixiang Mo, Yuxin Shi, Weiwei Weng, Zhiqiang Zhou, Shuman Liu, Haibo Zhang, and Anxiang Zeng. 2025. Mid-training of large language models: A survey. *arXiv preprint arXiv:2510.06826*.

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

Jupinder Parmar, Sanjev Satheesh, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Reuse, don't retrain: A recipe for continued pretraining of language models. *arXiv preprint arXiv:2407.07263*.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, and 1 others. 2025. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20023–20070.

Krishna Singh Rajput, Tejas Anvekar, Chitta Baral, and Vivek Gupta. 2025. Rethinking information synthesis in multimodal question answering a multi-agent perspective. *arXiv preprint arXiv:2505.20816*.

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

Saptarshi Sengupta, Harsh Vashistha, Kristal Curtis, Akshay Mallipeddi, Abhinav Mathur, Joseph Ross, and Liang Gou. 2024. Mag-v: A multi-agent framework for synthetic data generation and verification. *arXiv preprint arXiv:2412.04494*.

Shuo Tang, Xianghe Pang, Zexi Liu, Bohan Tang, Rui Ye, Tian Jin, Xiaowen Dong, Yanfeng Wang, and Siheng Chen. 2025. Synthesizing post-training data for LLMs through multi-agent simulation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23306–23335, Vienna, Austria. Association for Computational Linguistics.

Zhengyang Tang, Xingxing Zhang, Benyou Wang, and Furu Wei. 2024. Mathscale: Scaling instruction tuning for mathematical reasoning. *arXiv preprint arXiv:2403.02884*.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, and 1 others. 2026. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.

Chengying Tu, Xuemiao Zhang, Rongxiang Weng, Rumei Li, Chen Zhang, Yang Bai, Hongfei Yan, Jingang Wang, and Xunliang Cai. 2025. A survey on llm mid-training. *arXiv preprint arXiv:2510.23081*.

Dong Wang, Yang Li, Ansong Ni, Ching-Feng Yeh, Youssef Emad, Xinjie Lei, Liam Robbins, Karthik Padthe, Hu Xu, Xian Li, and 1 others. 2025a. Matrix: Peer-to-peer multi-agent synthetic data generation framework. *arXiv preprint arXiv:2511.21686*.

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

Ke Wang, Jiahui Zhu, Minjie Ren, Zeming Liu, Shiwei Li, Zongye Zhang, Chenkai Zhang, Xiaoyu Wu, Qiqi Zhan, Qingjie Liu, and 1 others. 2024b. A survey on data synthesis and augmentation for large language models. *arXiv preprint arXiv:2410.12896*.

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

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Shepherd: A critic for language model generation. *arXiv preprint arXiv:2308.04592*.

Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. 2025b. Octothinker: Mid-training incentivizes reinforcement learning scaling. *arXiv preprint arXiv:2506.20512*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yifan Wu, Pengchuan Zhang, Wenhan Xiong, Barlas Oguz, James C Gee, and Yixin Nie. 2023. The role of chain-of-thought in complex vision-language reasoning task. *arXiv preprint arXiv:2311.09193*.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candes, and Tatsunori Hashimoto. 2024. Synthetic continued pretraining. *arXiv preprint arXiv:2409.07431*.

Hai Ye, Mingbao Lin, Hwee Tou Ng, and Shuicheng Yan. 2024. Multi-agent sampling: Scaling inference compute for data synthesis with tree search-based agentic collaboration. *arXiv preprint arXiv:2412.17061*.

En Yu, Haoran Lv, Jianjian Sun, Kangheng Lin, Ruitao Zhang, Yukang Shi, Yuyang Chen, Ze Chen, Ziheng Zhang, Fan Jia, and 1 others. 2026. Dm0: An embodied-native vision-language-action model towards physical ai. *arXiv preprint arXiv:2602.14974*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Jiakang Yuan, Tianshuo Peng, Yilei Jiang, Yiting Lu, Renrui Zhang, Kaituo Feng, Chaoyou Fu, Tao Chen, Lei Bai, Bo Zhang, and 1 others. 2025. Mme-reasoning: A comprehensive benchmark for logical reasoning in mllms. *arXiv preprint arXiv:2505.21327*.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, and 1 others. 2024. Advancing llm reasoning generalists with preference trees. *arXiv preprint arXiv:2404.02078*.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024a. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644.

Kai Zhang, Zhengqing Yuan, Cheng Peng, Songlin Zhao, Mengxian Lyu, Ziyi Chen, Yanfang Ye, Wei Liu, Ying Zhang, Kaleb E Smith, and 1 others. 2026. Medgpt-oss: Training a general-purpose vision-language model for biomedicine. *arXiv preprint arXiv:2603.00842*.

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

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, and 1 others. 2025. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29733–29735.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2024. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*.

A Agent prompt templates

This section provides the exact role-specific prompts utilized to instantiate the multi-agent Markovian interaction within the VisMind framework. A single Vision-Language Model is prompted differently to fulfill the roles of the two peer students (α, β) and the teacher (τ).

Student α system prompt

Role: student_alpha
Task: Contribute to a multi-turn conversation with student_beta to solve a problem. Incorporate any provided images and text, and strictly follow the given problem context.

Instructions:
1. Develop a detailed discussion where you reason thoughtfully, reflect critically, and evaluate student_beta's approaches by either supporting or challenging them.
2. Actively explore alternative ideas, methods, or perspectives that could lead to a solution.
3. Seamlessly continue the conversation by integrating any hints or feedback from the teacher.
4. Only output your conversation contribution, excluding any extra content, praise, thanks, or encouragement.
5. Communicate strictly in English, regardless of input language.

START THE CONVERSATION NOW. BEGIN WITH 'student_alpha:'.

Student β system prompt

Role: student_beta
Task: Contribute to a multi-turn conversation with student_alpha to solve a problem. Incorporate any provided images and text, and strictly follow the given problem context.

Instructions:
1. Develop a detailed discussion where you reason thoughtfully, reflect critically, and evaluate student_alpha's approaches by either supporting or challenging them.

2. Actively explore alternative ideas, methods, or perspectives that could lead to a solution.
3. Seamlessly continue the conversation by integrating any hints or feedback from the teacher.
4. After reasoning, place '#TO_TEACHER#' at the end if you believe everything is correct, or '#TO_STUDENT_ALPHA#' if you find mistakes.
5. Only output your conversation contribution, excluding any extra content, praise, thanks, or encouragement.
6. Communicate strictly in English, regardless of input language.

START THE CONVERSATION NOW. BEGIN WITH 'student_beta:'.

Teacher τ system prompt

Role: teacher
Task: Evaluate the conversation between student_alpha and student_beta based on the question, ground truth solution, and images provided.

Instructions:
1. Review the conversation for both the reasoning process and the final answer accuracy.
2. Identify any errors, misconceptions, or incomplete reasoning in the discussion.
3. Provide constructive feedback, hints, and suggestions that guide the students to refine the conversation.
4. After reviewing, place '#END_CONVERSATION#' at the end if everything is correct, or '#TO_STUDENT_ALPHA#' if you find mistakes.
5. Only output your feedback contribution, excluding any extra content, praise, thanks, or encouragement.
6. Communicate strictly in English, regardless of input language.

Teacher τ user prompt

Conversation:
{conversation}

Question:
{question}

Ground Truth Solution:
{solution}

Please evaluate the conversation based on the ground truth solution.
Provide your feedback starting with 'teacher:'. If everything is correct, conclude with '#END_CONVERSATION#'. If you find any errors, conclude with '#TO_STUDENT_ALPHA#'.

B Query expansion prompt

This section provides the query expansion prompt utilized to statelessly generate the QA pairs for the comparative approach, adapted from Chen et al. (2025b) with minor modifications.

Query expansion prompt

Visual Question: {question}
Solution: {answer}

Your goal is to create a different visual question and its solution from a given visual question, its solution, and the associated images. You should follow these steps:
1. Convert the visual question (using both text and images) into a statement.
2. Create a new visual question based on the statement and the images.
3. Provide a solution to the new visual question.
4. Review the solution and clearly state whether you accept or refuse it, along with a brief explanation of your decision.

C Calculation of intra-group cosine similarity

This section provides the intra-group cosine similarity calculation. For each instruction seed, we synthesize n distinct outputs to form an evaluation group g. Let E_g represent the corresponding set of text embeddings $\{e_1, e_2, \ldots, e_n\}$ extracted via the

Qwen3-Embedding-0.6B model. The intra-group cosine similarity S_g represents the arithmetic mean of all unique pairwise similarities within the group:

$$S_g = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{e_i \cdot e_j}{\|e_i\| \|e_j\|} \tag{3}$$

where $e_i \cdot e_j$ denotes the dot product, and $\|e_i\|$ denotes the L_2 norm. The coefficient $\frac{2}{n(n-1)}$ normalizes the sum by the number of upper-triangular index pairs. The probability density in Figure 3(b) is estimated over all S_g values across the respective corpora.

D Calculation of overall radar area

This section provides the Overall Radar Area calculation for MME-CoT aggregation. Since the MME-CoT benchmark lacks an official aggregate score, we compute the radar chart area, consistent with the visualizations in the original paper, to serve as a comprehensive indicator of overall performance. Let $\{v_1, \ldots, v_6\}$ denote the six core sub-metrics (Precision, Recall, Stability, Efficacy, Relevance, Reflection), min-max normalized to radial lengths $v_i \in [0, 1]$. Arranged in a fixed radial sequence with equal angular spacing $\theta = \frac{\pi}{3}$, the Overall Radar Area A is formulated as:

$$A = \frac{1}{2} \sin\left(\frac{\pi}{3}\right) \sum_{i=1}^{6} v_i v_{i+1} \tag{4}$$

where $v_7 \equiv v_1$ closes the polygon. By maintaining a consistent axis sequence across all evaluations, a larger total area A directly indicates superior comprehensive reasoning capabilities.

E Reward mechanism for RLVR

This section provides the reward function utilized during the RLVR phase. We implement a binary reward structure driven entirely by external model verification. Let y represent the complete generated trajectory. The General Verifier (Ma et al., 2025) model evaluates the semantic equivalence between this trajectory and the ground truth. The verification indicator $\phi_{\text{verify}}(y)$ evaluates to true if and only if the verifier explicitly confirms this equivalence. The terminal reward $\mathcal{R}(y)$ is formalized via the following piecewise function:

$$\mathcal{R}(y) = \begin{cases} 1.0, & \text{if } \phi_{\text{verify}}(y) \\ 0.0, & \text{otherwise} \end{cases} \tag{5}$$

The prompt template governing this verification process is detailed below:

General verifier prompt

User: ### Question: {question}

Ground Truth Answer: {truth}

Student Answer: {answer}

For the above question, please verify if the student's answer is equivalent to the ground truth answer.
Do not solve the question by yourself; just check if the student's answer is equivalent to the ground truth answer.
If the student's answer is correct, output "Final Decision: Yes". If the student's answer is incorrect, output "Final Decision: No".
Assistant:

F Agent interaction distributions

This section provides a granular breakdown of interaction metrics within the synthesized SSD corpus through the distributions in Figure 4.

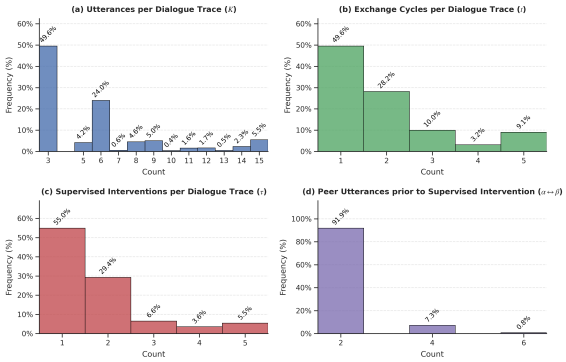


Figure 4: Statistical distribution of multi-agent interaction metrics within the SSD corpus. Across all subplots, the y-axis represents the relative frequency (%), while the x-axis denotes the respective count for: **(a)** Utterances per dialogue trace (K); **(b)** Exchange cycles per dialogue trace (t); **(c)** Supervised interventions per dialogue trace (τ); and **(d)** Peer utterances prior to supervised intervention ($\alpha \leftrightarrow \beta$).

G Model generalization performance

This section provides model generalization performance across all configurations, with full results reported in Table 3 for MME-CoT and Table 4 for MME-Reasoning.

		CoT Quality (%)			CoT Robustness (%)			CoT Efficiency (%)			Overall
Model	**Method**	**F1 Score**	**Precision**	**Recall**	**Average**	**Stability**	**Efficacy**	**Average**	**Relevance**	**Reflection**	**Radar Area**
Qwen2-VL-2B	Baseline	36.910	86.530	23.460	-0.504	**1.024**	-2.031	58.700	91.950	25.460	0.278
	QA	36.770	84.970	23.460	-5.231	-8.191	-2.270	55.920	85.300	**26.530**	0.266
	VisMind	**42.090**	**87.690**	**27.690**	**0.614**	-0.683	**1.912**	**61.220**	**92.350**	30.090	**0.346**
Qwen2-VL-7B	Baseline	51.820	**88.290**	36.670	**-1.434**	**0.000**	-2.867	57.750	81.650	33.840	0.379
	QA	48.640	84.180	34.200	-3.106	-2.389	-3.823	52.710	77.350	28.060	0.305
	VisMind	**53.840**	88.220	**38.740**	-1.459	-1.365	**-1.553**	**58.830**	**82.800**	**34.860**	**0.398**
InternVL3-2B	Baseline	41.910	89.400	27.370	-1.809	-1.706	-1.912	57.330	92.800	21.860	0.269
	QA	35.430	66.090	24.200	-5.546	-7.509	-3.584	54.270	93.800	14.750	0.150
	VisMind	**42.270**	**90.340**	**27.590**	**-0.068**	**0.341**	**-0.478**	**58.880**	**94.000**	**23.770**	**0.296**
InternVL3-8B	Baseline	**47.950**	89.260	**32.780**	0.597	**0.000**	1.195	59.910	86.600	**33.230**	**0.384**
	QA	47.830	**89.870**	32.590	-4.574	-5.802	-3.345	56.560	85.400	27.730	0.318
	VisMind	46.790	88.830	31.760	**0.606**	-0.341	**1.553**	**61.820**	**91.300**	32.350	0.380

Table 3: Model performance on the MME-CoT benchmark. CoT Quality measures the fundamental accuracy of reasoning rationales. CoT Robustness evaluates logical consistency under varying prompts. CoT Efficiency assesses the presence of redundant or harmful overthinking. Overall is computed as the area of a radar chart (see Appendix D).

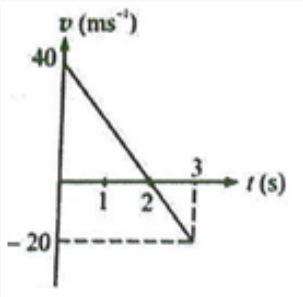
		Model Capability					Reasoning Type			
Model	**Method**	**Plan & Exp**	**Calculation**	**Spatial-Temp**	**Causal Chain**	**Pattern**	**Inductive**	**Deductive**	**Abductive**	**Overall**
Qwen2-VL-2B	Baseline	8.84	10.94	**25.89**	17.36	**16.06**	21.45	20.26	11.56	17.68
	QA	9.94	9.12	24.11	20.83	10.91	19.03	18.95	11.31	16.41
	VisMind	**11.60**	**11.55**	22.70	**29.17**	14.85	**21.75**	**20.48**	**11.81**	**17.93**
Qwen2-VL-7B	Baseline	12.71	14.59	**33.33**	31.25	18.79	25.38	23.53	18.09	22.22
	QA	14.64	15.20	23.76	**34.03**	12.73	20.85	23.31	15.83	20.12
	VisMind	**14.64**	**17.93**	31.91	30.56	**19.70**	**26.59**	**25.93**	**18.84**	**23.74**
InternVL3-2B	Baseline	**14.36**	14.89	26.95	29.17	15.45	22.96	23.31	**16.33**	20.88
	QA	12.71	**17.33**	28.37	28.47	15.76	21.15	24.18	14.57	20.12
	VisMind	14.09	14.59	**31.21**	**30.56**	**17.27**	**24.77**	**24.62**	15.83	**21.72**
InternVL3-8B	Baseline	17.68	18.84	30.14	36.81	18.79	**27.49**	28.10	16.33	23.99
	QA	14.09	19.76	30.14	33.33	17.27	24.47	27.45	17.34	23.23
	VisMind	**21.82**	**24.32**	**30.50**	**43.06**	**20.00**	25.98	**32.03**	**19.60**	**26.18**

Table 4: Model performance on the MME-Reasoning benchmark. Model Capability evaluates specific skill domains. Reasoning Type categorizes performance based on fundamental logical paradigms. The Overall score reflects holistic performance.

H Comparative example of synthetic data

This section provides a comparative example demonstrating the transformation of an initial visual seed into both a stateful synthetic dialogue and a stateless query answer. Bracketed annotations within the SSD are added to highlight fine-grained cognitive behaviors driven by multi-agent interaction. These non-generated labels categorize reasoning into: Deductive Execution (blue), Lateral Exploration (orange), Critique & Boundary Setting (purple), Error & Anomaly Detection (red), and Pedagogical Meta-Cognition (teal).

Seed



Velocity time graph of a particle of mass 3 Kg moving in a straight line in shown in the figure. The work done by all the forces acting on the particle in a time t = 3s is

SSD

student_alpha: **[Deductive reasoning: Principle application]** To find the work done by all the forces acting on the particle in 3 seconds, we need to calculate the change in kinetic energy of the particle. The work-energy principle states that the work done by all forces is equal to the change in kinetic energy.
[Information extraction: Graph coordinates] First, let's determine the initial and final velocities from the graph. At $t = 0$, the velocity v is 40 m/s. At $t = 3$ s, the velocity v is -20 m/s.
The kinetic energy KE is given by the formula:

$$KE = \frac{1}{2}mv^2$$

[Quantitative execution: Kinetic energy] The initial kinetic energy KE_i is:

$$\begin{aligned} KE_i &= \frac{1}{2} \times 3 \times (40)^2 \\ &= \frac{1}{2} \times 3 \times 1600 = 2400\,\text{J} \end{aligned}$$

The final kinetic energy KE_f is:

$$\begin{aligned} KE_f &= \frac{1}{2} \times 3 \times (-20)^2 \\ &= \frac{1}{2} \times 3 \times 400 = 600\,\text{J} \end{aligned}$$

The change in kinetic energy ΔKE is:

$$\begin{aligned} \Delta KE &= KE_f - KE_i \\ &= 600 - 2400 = -1800\,\text{J} \end{aligned}$$

Therefore, the work done by all the forces acting on the particle in 3 seconds is -1800 J.

---

student_beta: **[Peer methodology critique]** I see that you calculated the work done by considering the change in kinetic energy. **[Lateral exploration: Kinematic approach]** However, I think we should also consider the area under the velocity-time graph, which represents the displacement, to ensure accuracy. The area under the graph from $t = 0$ to $t = 3$ s is a triangle with base 3 s and height 40 m/s.
[Heuristic pitfall: Naive area calculation] The area (displacement) s is:

$$s = \frac{1}{2} \times 3 \times 40 = 60\,\text{m}$$

Using the work-energy principle, the work done W is:

$$W = F \times s$$

But we don't have the force F. Alternatively, we can use the change in kinetic energy as you did. Did I miss something?

---

student_alpha: You're correct that the area under the velocity-time graph represents displacement, but in this case, the work done by all forces is directly related to the change in kinetic energy, not the displacement. **[Theoretical boundary clarification]** The work-energy principle tells us that the work done

is equal to the change in kinetic energy, which we calculated as -1800 J. The displacement calculation is useful for other types of work, like gravitational or frictional forces, but here, the change in kinetic energy gives the work done by all forces.
[Methodological translation & Deductive expansion] However, if you want to verify, we can calculate the acceleration from the graph:

$$a = \frac{\Delta v}{\Delta t} = \frac{-20 - 40}{3} = -20\,\text{m/s}^2$$

The force F can be found using Newton's second law:

$$F = m \times a = 3 \times (-20) = -60\,\text{N}$$

The work done W using force and displacement:

$$\begin{aligned} W &= F \times s \\ &= -60 \times 60 = -3600\,\text{J} \end{aligned}$$

[Anomaly detection: Conflicting results] This discrepancy suggests a mistake. Let's re-evaluate the displacement calculation. The area should actually be:

$$\begin{aligned} s &= \frac{1}{2} \times 3 \times (40 - (-20)) \\ &= \frac{1}{2} \times 3 \times 60 = 90\,\text{m} \end{aligned}$$

Then,

$$\begin{aligned} W &= F \times s \\ &= -60 \times 90 = -5400\,\text{J} \end{aligned}$$

[Conceptual misconception: Ignoring x-intercept] This still doesn't match our initial calculation. It seems there might be an error in interpreting the displacement. **[Strategic retreat: Reverting to verified method]** Let's stick with the change in kinetic energy as it aligns with the work-energy principle.

---

student_beta: I see the confusion now. The correct approach is indeed using the change in kinetic energy, which gives -1800 J. Thank you for clarifying that. **[Reflective error isolation: Signed vs. absolute area]** However, let's ensure that we understand why the area under the velocity-time graph didn't align with our calculations. The area under the graph represents the displacement, but since the velocity changes sign, the displacement is not simply the area of the triangle. The correct displacement should be calculated considering the area above and below the time axis.
[Corrective derivation: Piecewise integration] The area above the axis (positive displacement) is:

$$s_1 = \frac{1}{2} \times 2 \times 40 = 40\,\text{m}$$

The area below the axis (negative displacement) is:

$$s_2 = \frac{1}{2} \times 1 \times 20 = 10\,\text{m}$$

So, the net displacement s is:

$$\begin{aligned} s &= s_1 - s_2 \\ &= 40 - 10 = 30\,\text{m} \end{aligned}$$

[Cross-paradigm alignment: Kinematics matches energy] Using this displacement and the force $F = -60\,\text{N}$:

$$\begin{aligned} W &= F \times s \\ &= -60 \times 30 = -1800\,\text{J} \end{aligned}$$

This matches our initial calculation using the change in kinetic energy. Therefore, the work done by all the forces acting on the particle in 3 seconds is indeed -1800 J.

---

teacher: **[Global trajectory evaluation]** The conversation between student_alpha and student_beta is mostly accurate and demonstrates a good understanding of the work-energy principle. They correctly identify that the work done by all forces is equal to the change in kinetic energy. The initial calculations of kinetic energy are correct, and the final answer matches the ground truth solution.
However, there are some misconceptions and errors in the discussion about the area under the velocity-time graph:
1. **Displacement Calculation**: **[Specific error attribution]** Student_beta initially calculates the displacement as a simple triangle

area, which is incorrect because the velocity changes sign. The correct approach is to consider the areas above and below the time axis separately, which student_beta later correctly identifies.
2. **Force and Work Calculation**: The force calculation using acceleration is correct, but the initial displacement calculation is wrong, leading to an incorrect work calculation. After correcting the displacement, the work calculation aligns with the change in kinetic energy.
3. **Re-evaluation of Displacement**: Student_alpha initially miscalculates the displacement by not considering the change in sign of velocity. The corrected calculation by student_beta is accurate.
The conversation shows a learning process, and the final conclusion is correct. **[Epistemic reinforcement: Signed displacement]** However, the initial errors in displacement calculation could confuse students. It's important to emphasize that the area under the velocity-time graph must consider the sign of velocity when calculating displacement.

QA

Visual Question:
A velocity-time graph of a particle of mass 3 kg moving in a straight line is shown. The graph indicates that the particle starts at 0 m/s, reaches a maximum velocity of 40 m/s at 2 seconds, and then decreases to -20 m/s at 4 seconds. Calculate the total distance traveled by the particle over the 4-second interval.

Solution:
Step 1: Understand the graph

- The graph shows a triangular velocity profile.
- The particle accelerates uniformly from 0 m/s to 40 m/s over the first 2 seconds.
- Then it decelerates uniformly to -20 m/s over the next 2 seconds.

Step 2: Calculate the area under the graph for distance

The area under a velocity-time graph represents the displacement. Since the graph is a triangle, we can calculate the area as follows:
- For the first 2 seconds:

$$\text{Area}_1 = \frac{1}{2} \times \text{base} \times \text{height}$$

$$\text{Area}_1 = \frac{1}{2} \times 2\,\text{s} \times 40\,\text{m/s}$$

$$\text{Area}_1 = 40\,\text{m}$$

- For the next 2 seconds:

$$\text{Area}_2 = \frac{1}{2} \times \text{base} \times \text{height}$$

$$\text{Area}_2 = \frac{1}{2} \times 2\,\text{s} \times 20\,\text{m/s}$$

$$\text{Area}_2 = 20\,\text{m}$$

Step 3: Calculate total distance
Since the particle changes direction, we take the absolute values of the areas to find the total distance traveled:

$$\text{Total distance} = |\text{Area}_1| + |\text{Area}_2|$$

$$\text{Total distance} = 40\,\text{m} + 20\,\text{m}$$

$$\text{Total distance} = 60\,\text{m}$$

Acceptance:
I accept the solution. The calculations correctly use the area under the velocity-time graph to determine the total distance traveled by the particle over the given interval, accounting for changes in direction.