Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Explain-on-Blackboard Paradigm for Interleaved Visual-Text Reasoning

Linzhuang Sun
Shenyang Institute of Computing Technology (China)
Sizhe Qiu
Shenyang Institute of Computing Technology (China)
Jianwei Cao
South China Normal University
Bihui Yu
Shenyang Institute of Computing Technology (China)
Zhaojin Fu
University of Science and Technology Beijing
Taishan Zeng
zengtsh@m.scnu.edu.cn

South China Normal University

Research Article

Keywords: Interleaved Visual-Text Reasoning, Knowledge Distillation, Multimodal Chain-of-Thought, Process-Oriented Fine-Tuning

Posted Date: May 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9433056/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

The Explain-on-Blackboard Paradigm for Interleaved Visual-Text Reasoning

Linzhuang Sun[1,2†], Sizhe Qiu[1,2†], Jianwei Cao[4†], Bihui Yu[1,2], Zhaojin Fu[3], Taishan Zeng[4*]

[1]Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang, 110168, China.
[2]University of Chinese Academy of Sciences, Beijing, 101408, China.
[3]University of Science and Technology Beijing, Beijing, 100083, China.
[4]South China Normal University, Guangzhou, 510631, China.

*Corresponding author(s). E-mail(s): zengtsh@m.scnu.edu.cn;
Contributing authors: sunlinzhuang21@mails.ucas.ac.cn;
qiusizhe25@mails.ucas.ac.cn; 20222232002@m.scnu.edu.cn;
yubihui@sict.ac.cn; D202410395@xs.ustb.edu.cn;
†These authors contributed equally to this work.

Abstract

Interleaved visual-text chain-of-thought reasoning, generating intermediate visual artifacts to support deduction, is a promising paradigm for complex multimodal tasks. However, models suffer a representation tax causing systematic performance degradation when translating abstract reasoning into structured spatial visualizations. We attribute this to alignment bias in conventional supervised fine-tuning, which optimizes labels while neglecting synchronization between visual semantics and textual reasoning trajectories. To address this, we propose a pedagogical distillation framework grounded in the Explain-on-Blackboard (EoB) paradigm. Guided by multimedia learning and social constructivism, our method requires teacher models to generate temporally and semantically aligned teaching trajectories comprising a discourse channel for logical explanation and a canvas channel for incremental HTML/SVG construction. A dynamic-rubric quality assurance mechanism with multi-stage filtering retains only high-fidelity trajectories for distillation. To validate our approach, we construct Cas-Bench, a 1,000-problem multidisciplinary protocol. Experiments show an 8B-parameter model fine-tuned on 4,000 distilled trajectories achieves a 69.76% relative improvement on Cas-Bench over baselines across MMMU, MMMU-Pro, MathVision, and MathVista benchmarks.

Keywords: Interleaved Visual-Text Reasoning, Knowledge Distillation, Multimodal Chain-of-Thought, Process-Oriented Fine-Tuning

1 Introduction

Interleaved visual-text chain-of-thought (CoT) reasoning, which uses procedural visual information to assist model inference, has become an important technique for solving complex multimodal problems [1, 2]. Recent works explore this direction from multiple angles. For example, Visual Sketchpad [3] generates process images through explicit Python-based plotting, whereas approaches like [4, 5] perform implicit visual information modeling within latent spaces. However, Cheng et al. [6] observe that naively combining procedural visual information with textual CoT can increase the reasoning burden of the model and lead to performance degradation. Although subsequent methods [7] use self-learning strategies to improve interleaved reasoning capabilities, a fundamental problem persists. During inference, the visual semantic features generated by current models exhibit a distributional misalignment with the accompanying textual semantics.

We argue that this alignment bias stems from a structural deficiency in conventional supervised fine-tuning (SFT). Standard SFT paradigms optimize predominantly toward outcome labels. They treat the final answer as the sole supervision signal and neglect explicit constraints on intermediate visual representations. Consequently, the attention of the model is not effectively guided to synchronize its visual generation with the evolving logical narrative. This observation motivates a core research question: how can we design a training methodology that enforces tight semantic and temporal alignment between textual reasoning and visual construction?

Beyond the training challenge, we identify a parallel deficiency in existing evaluation protocols. Current multimodal benchmarks, such as MMMU [8] and MathVista [9], rely mostly on outcome-based metrics that measure only final-answer accuracy. Such metrics inherently mask fragmented reasoning paths and cannot distinguish genuine visual-augmented reasoning from shortcut learning [10, 11]. A model that arrives at the correct answer has not necessarily reasoned coherently through a visual medium. This evaluative blind spot compounds the alignment problem, as models receive no gradient signal encouraging the quality of intermediate visual representations.

To address both challenges, we propose a pedagogical distillation framework based on discourse and canvas channels, grounded in the Explain-on-Blackboard" (EoB) paradigm. The core insight is drawn from cognitive science. In both human and artificial intelligence, pedagogical explanation represents a rigorous form of knowledge externalization [12]. Mayer's Cognitive Theory of Multimedia Learning and Paivio's Dual Coding Theory indicate that deep comprehension requires the simultaneous synchronization of verbal and visual channels [13, 14]. Viewed through the lens of Vygotsky's Zone of Proximal Development [15], the blackboard" serves as a cognitive scaffold that forces the model to externalize intermediate abstractions and construct the spatial structure necessary to convey its logic. This requirement imposes stronger

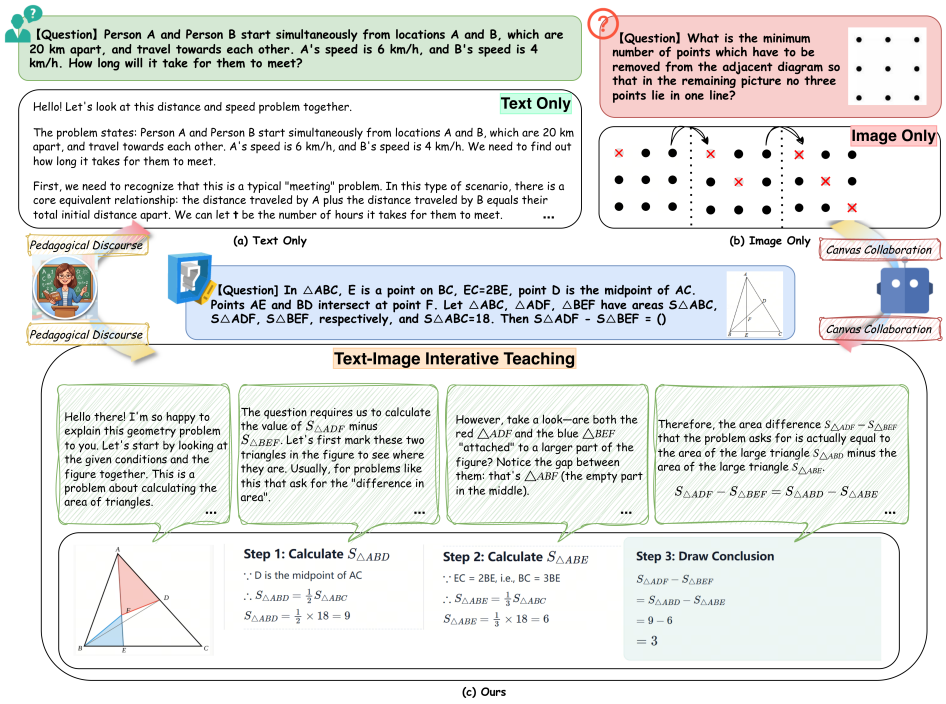


Fig. 1: Overview of the proposed dual-stream pedagogical distillation framework. Given a problem input, a teacher model generates synchronized Pedagogical Discourse (textual logic) and Canvas Collaboration (incremental HTML/SVG visuals), simulating a real-time blackboard teaching process. The distilled trajectories are used to fine-tune lightweight student models.

cognitive constraints than answer-seeking alone, as it mandates decomposing internal reasoning into a sequence of transparent, spatially grounded, and verifiable steps.

Our methodology operates in three stages. First, we use frontier teacher models (Gemini-3-Pro and Claude-Opus-4.5) to generate teaching trajectories comprising a discourse channel for step-by-step textual explanation and a canvas channel for incremental HTML/SVG visual construction, maintaining strict temporal and semantic alignment between the two. Second, we develop a Dynamic-Rubric quality assurance mechanism that automatically translates abstract pedagogical dimensions into concrete, domain-specific, machine-verifiable rubric items, enabling scalable multi-stage filtering of the generated trajectories. Third, we perform process-oriented SFT on lightweight student models using the curated teaching trajectory data, providing supervision signals that reward the alignment of logical deduction with visual representation.

To validate our approach, we construct Cas-Bench, an evaluation protocol comprising 1,000 expert-curated problems from three authoritative sources (CoMT [6], MathVision [16], and MMMU-Pro [17]) that span mathematics, natural sciences,

engineering, and humanities. Our empirical investigation reveals that frontier models suffer from a "representation tax". For instance, Gemini-3-Pro achieves 73.72 on the discourse channel but drops to 66.31 on the canvas channel, which quantifies the cognitive overhead of translating abstract logic into structured code. Fine-tuning an 8B-parameter model on 4,000 distilled trajectories yields a 69.76% relative improvement on Cas-Bench and absolute gains of 3 to 4 points across MMMU, MMMU-Pro, MathVision, and MathVista.

In summary, our main contributions are:

- We propose a pedagogical distillation framework using discourse and canvas channels that enforces temporal and semantic alignment between textual reasoning and visual construction, providing a principled training methodology for interleaved visual-text reasoning.
- We formalize the "Explain-on-Blackboard" paradigm based on cognitive and pedagogical theories from Mayer, Paivio, and Vygotsky and develop a dynamic-rubric quality assurance mechanism for scalable data curation.
- We introduce Cas-Bench, a 1,000-problem evaluation protocol with process-oriented metrics, to serve as a diagnostic tool for assessing interleaved reasoning capabilities.
- We demonstrate that process-oriented fine-tuning on the distilled teaching trajectories yields improvements on Cas-Bench and four established multimodal benchmarks, confirming the effectiveness and generalizability of the proposed methodology.

2 Related Work

2.1 From Linguistic CoT to Visual Externalization

CoT prompting has improved the deductive capabilities of language models by encouraging step-by-step linguistic reasoning [18, 19]. However, as tasks increase in spatial and high-dimensional complexity [20, 21], purely linguistic CoT often proves insufficient. To address this, recent efforts explore interleaved reasoning and tool-augmented generation, which allow models to invoke external visual modules or generate intermediate diagrams [22–24]. Notably, Canvas-CoT [25] proposes using HTML/SVG as a universal substrate for visual thought. Although theoretically promising, executing such visual reasoning imposes a substantial "representation tax", requiring both robust logical deduction and precise structured code synthesis. Our work addresses this bottleneck by proposing a distillation-based training methodology that systematically teaches models to align visual construction with textual reasoning.

2.2 Knowledge Distillation for Multimodal Reasoning

Knowledge distillation from frontier LMMs is standard practice for training lightweight open-source models [26, 27]. Traditional instruction-tuning datasets focus primarily on visual question answering or generic conversational alignment, emphasizing the mapping between an input image and a correct textual response [28–30]. While some recent works [31] attempt to distill reasoning traces, they rarely enforce strict temporal and semantic alignment between textual explanations and corresponding visual

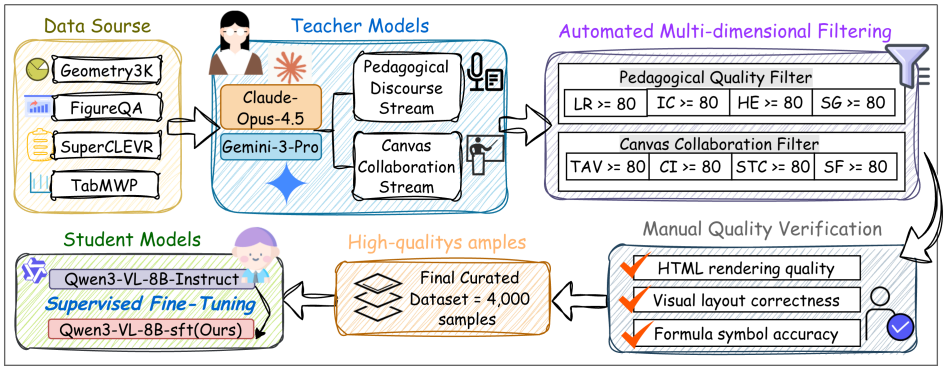


Fig. 2: The teaching trajectories distillation pipeline. Teacher models generate synchronized pedagogical trajectories, which undergo dynamic-rubric quality assurance and multi-stage filtering before being used to fine-tune student models.

updates. Our approach departs from standard VQA-style synthesis by explicitly modeling teaching trajectories composed of discourse and canvas channels. We distill not only the answer but the teaching-oriented construction process, ensuring that fine-tuned student models learn to synchronize logical discourse with executable canvas manipulation.

2.3 Multimodal Reasoning Benchmarks

The rapid development of LMMs has driven the creation of various benchmarks for complex reasoning [32, 33]. Datasets such as MMMU [8], MathVision [16], and CoMT [6] establish rigorous testing grounds for multimodal comprehension. However, the dominant evaluation paradigm relies on end-task accuracy [10, 11], which conflates genuine deductive reasoning with coincidental correctness. Cas-Bench complements these benchmarks by providing a process-oriented evaluation protocol that assesses the structural and semantic quality of intermediate visual artifacts, serving as a diagnostic tool to validate our proposed methodology.

3 Methodology

This section outlines the methodology for distilling teaching trajectories. We define the problem and the representation tax phenomenon (Section 3.1), establish the theoretical foundation of the Explain-on-Blackboard (EoB) paradigm (Section 3.2), and detail the generation protocol (Section 3.3). We then describe the dynamic-rubric quality assurance mechanism (Section 3.4) and the process-oriented fine-tuning strategy (Section 3.5).

3.1 Problem Formulation: The Representation Tax

We model the generation process of a Large Multimodal Model (LMM) as follows. Given a problem $P = (t, v)$ where t denotes the textual query and v represents an optional visual input, the model generates a latent reasoning chain $Z = (z_1, z_2, \ldots, z_T)$, where each z_t represents a discrete reasoning step. This latent process is externalized into two observable channels.

The discourse channel, serving as pedagogical discourse (PD), provides a textual explanation $D = (d_1, \ldots, d_T)$ where each segment d_t is a linguistic manifestation of latent step z_t, modeled by the emission probability $p(D \mid Z)$.

The canvas channel, serving as canvas collaboration (CC), consists of a sequence of manipulation commands $C = (c_1, \ldots, c_T)$ where each command c_t is a structured HTML/SVG representation of z_t, modeled by $p(C \mid Z)$.

The ideal interleaved reasoning process requires both channels to be high-quality and temporally synchronized. We formalize the core bottleneck as the representation tax:

$$Tax_{rep} = Q(S_{PD}) - Q(S_{CC}), \tag{1}$$

where $Q(\cdot)$ denotes a quality function over the respective channel. A high Tax_{rep} indicates that a model possesses strong linguistic deductive capabilities but struggles to map these abstractions onto structured spatial representations. This tax arises from the necessity to simultaneously maintain logical state, adhere to strict code syntax, and perform precise spatial planning.

This representation tax is not a mere capability gap to be measured, but a training deficiency to be addressed. Conventional SFT optimizes $p(\text{answer} \mid P)$, providing no gradient signal for the quality of intermediate visual representations. We optimize a process-oriented objective that supervises both channels jointly:

$$\mathcal{L} = -\sum_{t=1}^{T} \left[\log p(d_t \mid d_{<t}, c_{<t}, P) + \log p(c_t \mid d_{\leq t}, c_{<t}, P)\right], \tag{2}$$

The cross-channel conditioning ensures that the textual and visual outputs are generated in a mutually informed, temporally aligned manner.

3.2 The Explain-on-Blackboard Framework

To instantiate the process-oriented training objective, we require high-quality training data where textual reasoning and visual construction are tightly synchronized. The Explain-on-Blackboard paradigm provides the cognitive framework for generating such data.

This paradigm transitions from simple problem-solving to pedagogical explanation, as teaching represents a rigorous form of knowledge externalization [12]. The framework rests on three complementary cognitive theories.

Dual coding theory [14] posits that human cognition operates through functionally independent but interconnected verbal and non-verbal channels. Deep comprehension arises when information is encoded simultaneously in both. In our framework,

this translates to the requirement that every reasoning step must produce a textual explanation and a corresponding visual update.

The cognitive theory of multimedia learning [13] extends dual coding to instructional design, establishing that meaningful learning occurs when learners actively select, organize, and integrate verbal and pictorial representations. This motivates our requirement for temporal synchronization, ensuring visual elements appear in lockstep with their corresponding textual explanations rather than as post-hoc summaries.

Vygotsky's zone of proximal development [15] provides the scaffolding rationale. The blackboard serves as a cognitive scaffold that forces the model to externalize intermediate abstractions, architect the spatial structure necessary to convey its logic, and build complexity incrementally. This requirement eliminates shortcut learning, as the model cannot skip intermediate steps without producing visible gaps in the visual narrative.

We select HTML/SVG as the universal substrate for visual thought. Compared to bitmap image generation, its syntactic hierarchy imposes structural precision that mirrors logical reasoning structure. It is natively interpretable, allowing direct inspection of the visual reasoning process, and supports incremental construction through DOM manipulation to naturally align with step-by-step pedagogical explanation.

3.3 Teaching Trajectories Distillation

The distillation pipeline comprises teacher-driven generation and multi-stage quality filtering.

3.3.1 Teacher-Driven Generation

We use Gemini-3-Pro and Claude-Opus-4.5 as teacher models to synthesize pedagogical demonstrations. Given a problem instance from Geometry3K [34], FigureQA [35], SuperCLEVR [36], or TabMWP [37], the teacher is prompted to act as a virtual tutor to generate coherent teaching trajectories.

The discourse channel provides a step-by-step linguistic explanation that articulates the logical deduction process, supplying conceptual grounding and explicit reasoning transitions. Concurrently, the canvas channel produces a sequence of executable HTML/SVG operations that incrementally build an interpretable visual scaffold synchronized with the discourse.

Our generation protocol enforces strict temporal and semantic alignment between these channels. Each reasoning step in the discourse must trigger a meaningful update on the visual canvas, transforming the complex task of teaching via a blackboard into a structured supervision signal. Using two distinct teacher models also mitigates teacher-specific bias in the distilled data.

3.3.2 Multi-Stage Quality Filtering

Raw teacher outputs vary in quality. To guarantee that only high-fidelity trajectories enter the training set, we implement a rigorous filtering pipeline consisting of automated evaluation and human verification.

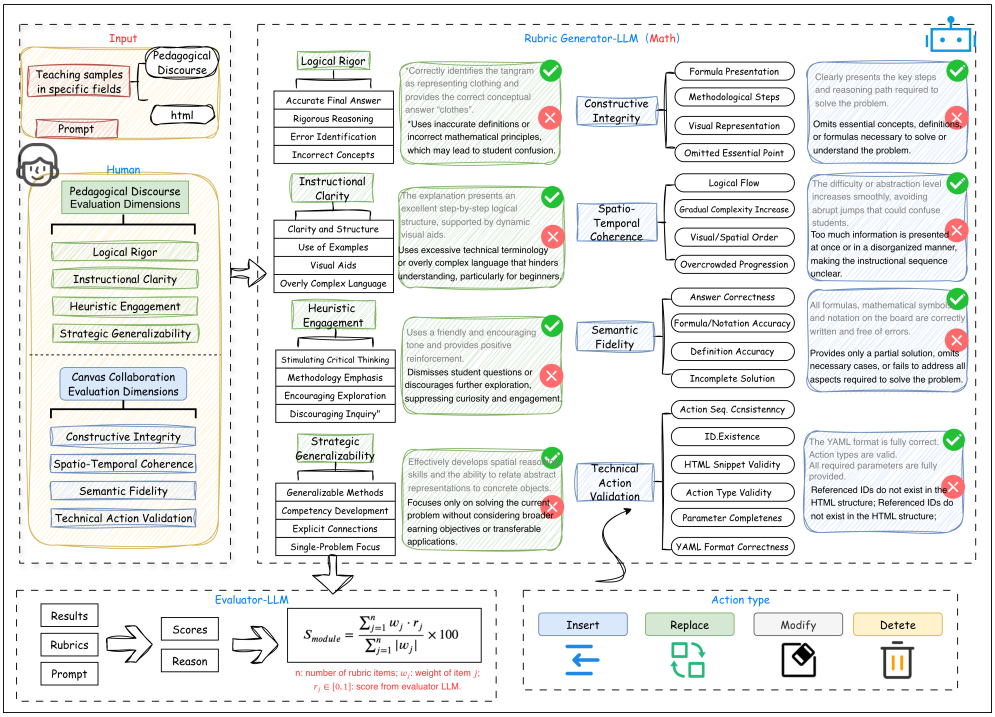


Fig. 3: The dynamic-rubric quality assurance mechanism. Expert-defined pedagogical dimensions are automatically expanded into domain-specific, machine-verifiable rubric items with differentiated weights, including essential ($w = 3$), important ($w = 2$), optional ($w = 1$), and penalty-based pitfall items ($w = -1$).

Each generated trajectory is automatically reviewed using a dynamic-rubric mechanism, detailed in Section 3.4. We use a high-capability reasoning model as the evaluator and impose strict admission thresholds. For the discourse channel, each trajectory is scored across logical rigor, instructional clarity, heuristic engagement, and strategic generalizability. A trajectory is accepted only if the score in each dimension exceeds 80 out of 100 and no incorrect concepts penalty is triggered. For the canvas channel, a formal code verification engine enforces 100% compliance with technical action validation. The generated canvas must also achieve an automated score of at least 75 in both constructive integrity and spatio-temporal coherence.

Following automated filtering, human annotators perform a comprehensive review of the admitted samples. They focus on subtle issues that automated evaluation might overlook, such as whether the HTML/SVG-rendered graphical layout is visually well-structured and whether mathematical symbols contain recognition errors. Through this mechanism, we retained approximately 4,000 high-quality data samples from the initial pool.

3.4 Dynamic-Rubric Quality Assurance

Assessing trajectory quality during the filtering stage requires a robust evaluation mechanism. Manually designing detailed rubric items for every domain is unscalable, so we address this through an automated rubric construction process.

3.4.1 Expert Framework Formulation

The foundation of our quality assurance system is an evaluation framework established by a committee of twelve professors from diverse academic fields. The framework decomposes teaching quality into dimensions grounded in educational psychology and visual communication principles for both the discourse and canvas channels.

The discourse channel dimensions comprise logical rigor, instructional clarity, heuristic engagement, and strategic generalizability. Logical rigor assesses whether the model presents a valid chain of deductive reasoning without gaps or fallacies. Instructional clarity evaluates the model's ability to structure its explanation cleanly, decompose complex concepts, and use accessible language. Heuristic engagement assesses whether the model encourages critical thinking by posing reflective questions. Strategic generalizability evaluates whether the model abstracts the problem-solving process into a strategy applicable to a broader class of problems.

The canvas channel dimensions include constructive integrity, spatio-temporal coherence, semantic fidelity, and technical action validation. Constructive integrity checks if the canvas includes all information required to understand the solution. Spatio-temporal coherence evaluates whether the arrangement and sequence of visual elements align with the logical flow of the explanation. Semantic fidelity verifies the correctness of visual elements like mathematical notations and labels. Technical action validation uses a programmatic engine to check HTML/SVG syntax validity, action types, and parameter completeness.

3.4.2 LLM-Driven Rubric Generation

Because the expert-defined dimensions are too abstract for direct use in an automated scoring system, we use a language model to translate them into concrete, domain-specific rubric items through a three-step process.

First, we provide the model with the dimension's formal definition, the specific academic domain of the task, and few-shot examples showing high-quality trajectories within that domain. Second, the model generates four verifiable sub-items that instantiate the abstract dimension in context. For instance, when processing logical rigor for a math problem, the model generates criteria such as ensuring the final answer is arithmetically correct and each derivation step logically follows from the previous one. Third, each generated sub-item is assigned a weight reflecting its pedagogical importance:

$$w_j = \begin{cases} 3, & \text{essential,} \\ 2, & \text{important,} \\ 1, & \text{optional,} \\ -1, & \text{pitfall,} \end{cases} \tag{3}$$

The inclusion of pitfall items with negative weights is a critical design choice. These items describe common teaching mistakes, such as using overly complex jargon, which introduces a non-monotonic penalty to ensure that superficial fluency cannot mask underlying logical contradictions.

3.4.3 Automated Scoring

An evaluator model assesses the response against each rubric item to assign a normalized score $r_j \in [0, 1]$. The quality score for each channel is computed as:

$$Q(\text{Channel}) = \frac{\sum_{j=1}^{m} w_j \cdot r_j}{\sum_{j=1}^{m} |w_j|} \times 100, \tag{4}$$

where m denotes the total number of rubric items and $|w_j|$ normalizes the score. The Cas-Bench Composite Score (CBS) combines both channels:

$$\text{CBS} = \frac{Q_{\text{PD}} + Q_{\text{CC}}}{2}. \tag{5}$$

This hierarchical scoring provides the quality threshold for data filtering and forms the evaluation metric for benchmarking.

3.4.4 Evaluation Consistency

Traditional evaluation approaches often suffer from high variance in complex reasoning tasks. By decomposing the evaluation into a hierarchical model with bounded, normalized sub-items, we mathematically constrain the variance of the estimator. Empirical validation against human expert judgments yields a Pearson correlation exceeding 0.94 across all eight evaluation dimensions.

3.5 Process-Oriented Fine-Tuning

With the curated dataset, we perform supervised fine-tuning on lightweight student models. The training objective follows Eq. (2), where the model learns to generate interleaved discourse and canvas tokens conditioned on the problem input and preceding context from both channels.

To isolate the contributions of our specific data structure, we design three training variants for ablation. In the text-only variant, the student is fine-tuned exclusively on the discourse channel, discarding all visual canvas operations. In the image-only variant, the student is fine-tuned solely on the final visual canvas artifacts, without the accompanying logical explanations. In our proposed teaching trajectories approach, the student is fine-tuned on the complete, synchronized dataset where textual discourse and incremental canvas construction are tightly aligned.

This ablation design directly tests the hypothesis that the alignment between textual reasoning and visual construction, rather than either modality in isolation, is the critical factor for effective interleaved reasoning.

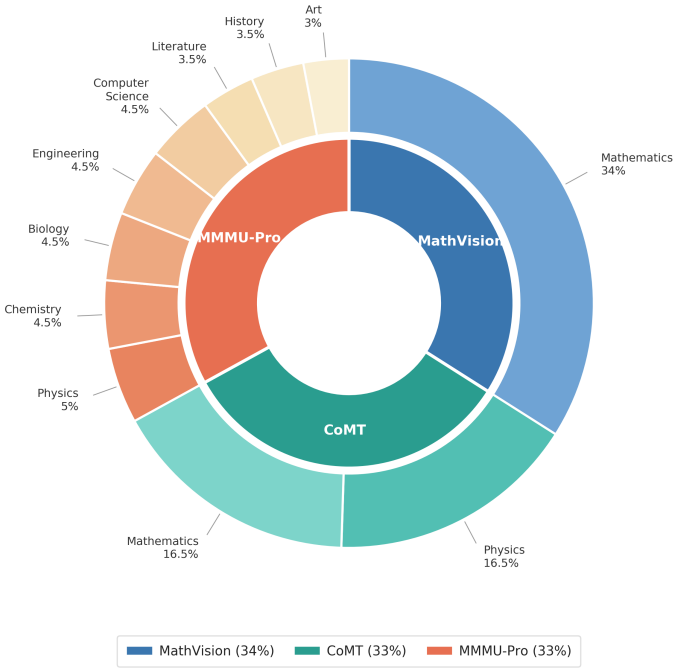


Fig. 4: The distribution of the 1,000 benchmark tasks in Cas-Bench. The inner ring represents the source datasets, and the outer ring specifies the corresponding academic disciplines.

4 Evaluation Protocol: Cas-Bench

We construct Cas-Bench, a diagnostic benchmark designed to evaluate interleaved visual-text reasoning.

4.1 Benchmark Construction

Cas-Bench comprises 1,000 problems sourced from three multimodal benchmarks: CoMT [6], MathVision [16], and MMMU-Pro [17]. The problems are distributed across multiple academic disciplines, including mathematics (approximately 40%), natural sciences (30%), engineering and technology (15%), and humanities and arts (15%). This distribution ensures that the benchmark tests both mathematical problem-solving and the ability to adapt visual representations to the characteristics of different fields. The overall statistical information is shown in Figure 4.

As illustrated in Figure 5, each problem is defined as a pair (t, v), where t denotes the textual query and v represents an optional visual input. The model is prompted to produce teaching trajectories consisting of a discourse channel and a canvas channel, forming a synchronized, interleaved process of explanation and visualization.

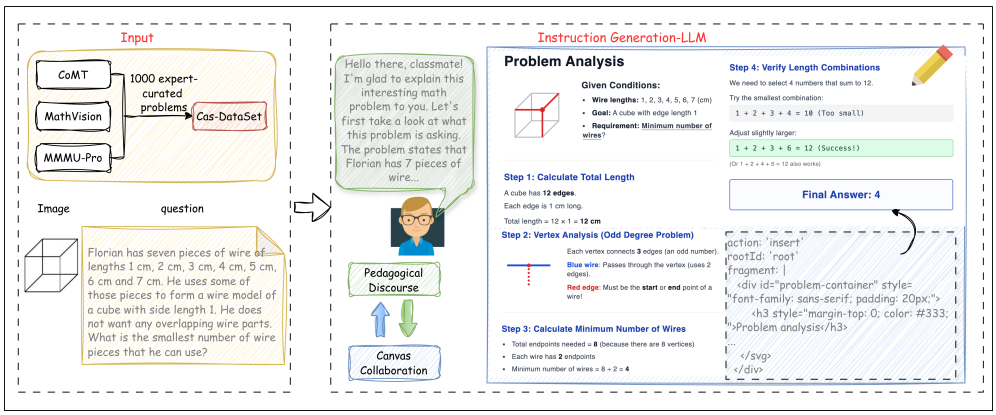


Fig. 5: The input and output structure of the Cas-Bench task. The pipeline illustrates the transformation of multi-domain problems into synchronized teaching trajectories comprising discourse and canvas channels.

4.2 Evaluation Metrics

The evaluation follows the dynamic-rubric framework described in Section 3.4. Each model output is scored across eight dimensions, with four allocated to each channel. The CBS metric aggregates these into a single composite score. The representation tax $Tax_{rep} = Q(S_{PD}) - Q(S_{CC})$ quantifies the performance gap between linguistic reasoning and visual externalization for each model.

5 Experiments

5.1 Experimental Setup

We conduct a comprehensive evaluation on Cas-Bench and four established multimodal benchmarks: MMMU (val) [8], MMMU-Pro [17], MathVision [16], and MathVista (testmini) [9]. The evaluated models span 16 systems across two categories: (1) proprietary frontier models, including Gemini-3-Pro, GPT-5.2, GPT-4o, Gemini-2.5-Pro, Claude-Opus-4.5, Claude-Sonnet-4, and Claude-3-Sonnet; and (2) open-source models of various scales, including Qwen3-VL-235B, InternVL2.5-78B, LLaVA-OneVision-72B, Qwen2.5-VL-72B, Yi-VL-34B, Phi-4-Vision-14B, Qwen3-VL-8B, and Qwen2.5-VL-7B. Our fine-tuned models (Qwen3-8B-sft and Qwen3-4B-sft) use Qwen3-VL-8B-Instruct and Qwen3-VL-4B-Instruct as base models, respectively. The sampling temperature is fixed at 0.0 for all models to ensure reproducibility.

5.2 The Representation Tax in Frontier Models

Table 1 and Figure 6 reveal a performance stratification across 16 models that supports our hypothesis regarding the representation tax. Proprietary models establish the upper bound, with Gemini-3-Pro leading at 70.02, followed by GPT-5.2 (68.15) and

Table 1: Main evaluation results on Cas-Bench. *Final* denotes the arithmetic mean of PD and CC channel scores. All scores are on a scale of [-100, 100]. Tax denotes the representation tax ($Q_{PD} - Q_{CC}$). The best score in each column is in **bold**, and the second best is underlined.

			Pedagogical Discourse (PD)					Canvas Collaboration (CC)				
Model	**Final**	Tax	**Score**	**LR**	**IC**	**HE**	**SG**	**Score**	**CI**	**STC**	**SF**	**TAV**
Proprietary Models												
Gemini-3-Pro	**70.02**	7.41	**73.72**	**75.14**	**77.89**	**72.54**	**69.28**	**66.31**	**49.06**	64.82	**67.25**	84.10
GPT-5.2	68.15	4.70	70.50	72.30	75.10	68.20	66.40	65.80	47.50	66.10	66.50	83.10
Gemini-2.5-Pro	65.65	2.89	67.10	69.58	73.81	62.79	62.16	64.21	46.67	65.17	64.97	80.02
Claude-Opus-4.5	62.77	**-4.71**	60.41	57.98	65.53	63.72	54.38	65.12	44.88	**67.27**	63.79	**84.52**
Claude-Sonnet-4	61.30	3.70	62.70	60.45	67.80	63.25	59.30	59.00	39.50	61.85	54.90	79.75
GPT-4o	60.25	6.50	63.50	61.20	68.45	62.30	62.05	57.00	37.85	58.40	52.75	79.00
Claude-3-Sonnet	55.43	2.06	56.46	55.10	61.80	52.50	56.45	54.40	35.70	58.10	43.30	80.50
Open-Source Models												
Qwen3-VL-235B	49.78	4.30	51.93	42.35	59.47	59.21	46.64	47.63	25.12	44.53	37.62	83.26
InternVL2.5-78B	45.30	6.70	48.65	39.80	55.40	51.20	48.20	41.95	19.45	39.70	27.85	80.80
Qwen3-8B-sft (Ours)	42.55	6.50	45.80	35.50	51.20	48.10	38.40	39.30	15.10	35.80	25.10	81.20
LLaVA-OneVision-72B	37.25	5.70	40.10	29.50	48.30	43.10	39.50	34.40	9.80	30.20	19.60	78.00
Qwen3-4B-sft (Ours)	35.80	6.20	38.90	28.50	46.20	41.60	39.30	32.70	11.20	29.40	18.50	71.70
Yi-VL-34B	34.65	6.14	37.72	27.90	44.80	40.50	37.70	31.58	8.50	28.40	18.62	70.80
Phi-4-Vision-14B	31.40	7.60	35.20	23.80	43.50	37.40	36.10	27.60	4.20	23.50	13.70	69.00
Qwen2.5-VL-72B	28.43	8.55	32.70	16.99	45.23	44.61	23.95	24.15	-1.95	16.56	4.66	77.33
Qwen3-VL-8B	27.04	7.78	30.93	12.30	40.11	44.31	25.63	23.15	1.02	15.51	5.34	70.72
Qwen2.5-VL-7B	20.33	3.05	21.85	14.47	34.88	23.67	12.86	18.80	-4.06	14.43	2.09	62.78

Gemini-2.5-Pro (65.65). Native open-source models struggle by comparison; the 235B-parameter Qwen3-VL scores 49.78, while the 72B-parameter Qwen2.5-VL reaches only 28.43. This disparity indicates that current open-source models cannot autonomously architect the spatial visual scaffolds required for interleaved reasoning.

Almost all models systematically score higher on PD than on CC, as shown in Figure 6. The representation tax is pervasive. Qwen2.5-VL-72B exhibits the largest gap ($Tax = 8.55$), followed by Qwen3-VL-8B ($Tax = 7.78$) and Phi-4-Vision-14B ($Tax = 7.60$). Open-source models frequently output negative scores in Constructive Integrity (e.g., Qwen2.5-VL-7B at -4.06, Qwen2.5-VL-72B at -1.95), reflecting deficiencies in generating critical visual components.

Claude-Opus-4.5 provides a notable exception by exhibiting a negative representation tax ($Tax = -4.71$) and achieving a higher CC score (65.12) than PD score (60.41). This model excels in Spatio-Temporal Coherence (67.27, the highest among all models) and Technical Action Validation (84.52), suggesting a strong capability to orchestrate progressive visual construction despite weaker linguistic pedagogy. This divergence implies that the two channels require independent optimization and that the representation tax is not strictly tied to overall model capacity.

Figure 7(a) further shows that model scale alone fails to resolve the interleaved reasoning challenge. The 72B-parameter LLaVA-OneVision (37.25) and Qwen2.5-VL-72B (28.43) underperform our fine-tuned 8B model (42.55), demonstrating that process-oriented supervision compensates for parameter disadvantages. Figure 7(b) illustrates that the representation tax remains consistent across model families. Open-source base models average $Tax \approx 6.2$ compared to $Tax \approx 3.2$ for proprietary models, indicating that proprietary models likely benefit from more balanced cross-modal training.

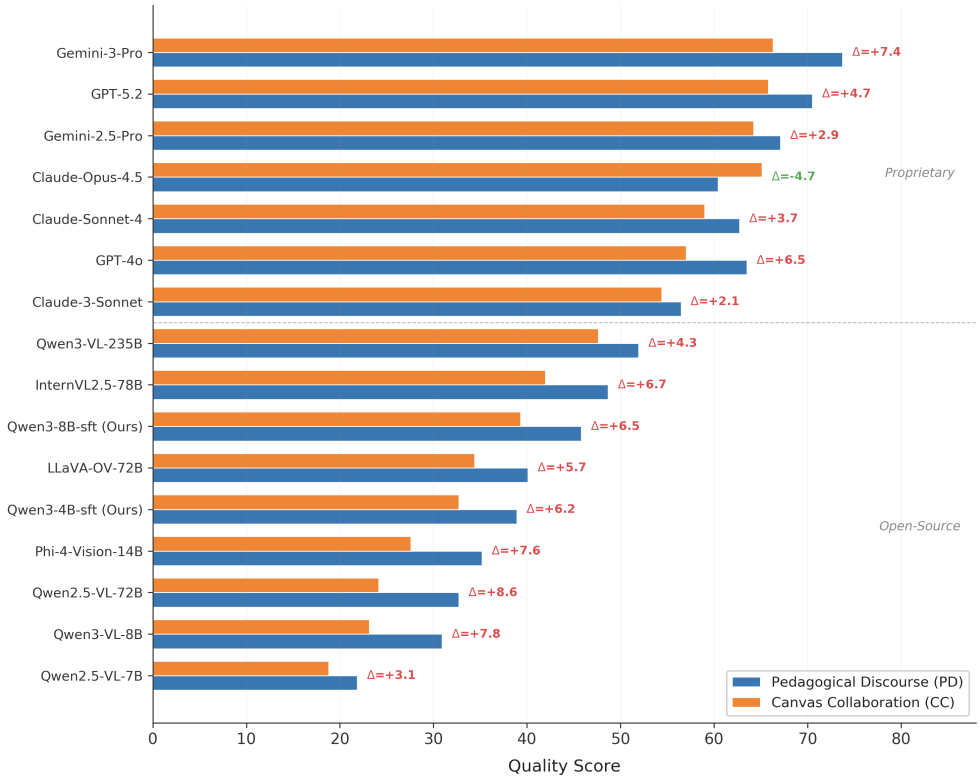


Fig. 6: Visualization of the representation tax across all 16 evaluated models. Blue bars represent Pedagogical Discourse (PD) scores; orange bars represent Canvas Collaboration (CC) scores. The annotated values above each pair indicate the representation tax ($Tax = Q_{PD} - Q_{CC}$). A positive tax (red) indicates that linguistic reasoning outperforms visual construction; a negative tax (green) indicates the reverse.

5.3 Effectiveness of Distilling Teaching Trajectories

Table 2 validates our methodology, showing that distilling teaching trajectories provides the highest quality supervision across all benchmarks, student capacities (4B/8B), and teacher models.

Channel alignment is the critical factor.

Fine-tuning the Qwen3-VL-8B model with Gemini-sourced teaching trajectories achieves substantial absolute gains over the baseline, reaching 73.2 on MMMU (+3.6) and 80.6 on MathVista (+3.4). These improvements remain consistent across both teacher models, ruling out teacher-specific artifacts.

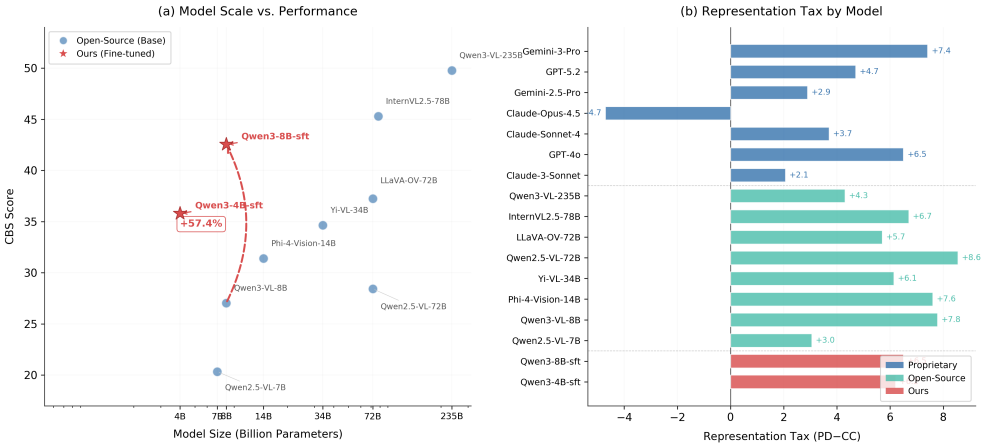


Fig. 7: (a) Relationship between model scale and CBS performance for open-source models. Our fine-tuned 8B model (red star) surpasses models with 4–9× more parameters. The dashed arrow shows the improvement from the 8B base to the fine-tuned version (+57.4%). (b) Representation tax by model, showing that the PD-CC gap is pervasive across both proprietary and open-source models.

Table 2: Ablation study on the effectiveness of distilling teaching trajectories. We compare our approach against text-only and image-only baselines using two student models and two teacher models across four established benchmarks.

Teacher Model	Training Data	MMMU (val)	MMMU-Pro	MathVision	MathVista
Student Model: Qwen3-VL-4B-Instruct					
No Teacher	-	67.4	53.2	51.6	73.7
Gemini-3-pro	Text only	68.6	54.5	53.1	75.2
	Image only	67.2	53.4	51.5	73.8
	Ours	**70.8**	**56.5**	**55.2**	**77.1**
Claude-Opus-4.5	Text only	69.1	54.8	52.9	74.9
	Image only	67.5	53.1	51.8	73.5
	Ours	**71.1**	**57.0**	**54.8**	**76.8**
Student Model: Qwen3-VL-8B-Instruct					
No Teacher	-	69.6	55.9	53.9	77.2
Gemini-3-pro	Text only	71.2	57.1	55.4	78.8
	Image only	69.4	56.1	53.8	77.0
	Ours	**73.2**	**59.4**	**57.5**	**80.6**
Claude-Opus-4.5	Text only	70.8	57.5	55.1	79.1
	Image only	69.7	55.7	54.1	77.3
	Ours	**73.5**	**59.8**	**57.2**	**81.1**

Unimodal training is insufficient.

The "Image only" configuration yields negligible or negative improvements (dropping the 4B baseline from 67.4 to 67.2 on MMMU), confirming that isolated visual artifacts lack actionable deductive signals without a logical narrative. "Text only" supervision offers moderate gains through conventional CoT distillation but consistently underperforms our aligned approach by 2–3 points. Purely linguistic CoT fails to ground complex spatial logic.

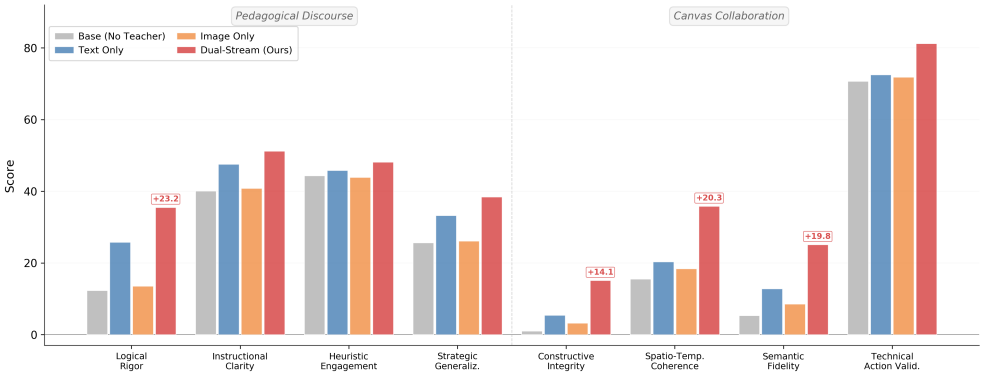


Fig. 8: Per-dimension ablation of training data variants on the Qwen3-VL-8B student model. The teaching trajectories approach (red) yields the largest gains in Canvas Collaboration dimensions (CI, STC, SF), confirming that the alignment between text and visuals, rather than either modality alone, is the critical supervision signal. Numerical annotations show the absolute gain over the base model for key dimensions.

Cas-Bench performance.

The fine-tuned Qwen3-8B-sft achieves a Final Score of 42.55, representing a 69.76% relative improvement over the base Qwen3-VL-8B (27.04). The smaller Qwen3-4B-sft reaches 35.80, surpassing the 34B-parameter Yi-VL (34.65) and approaching the 72B-parameter LLaVA-OneVision (37.25). The CC score of Qwen3-8B-sft (39.30) exceeds that of the much larger Qwen2.5-VL-72B (24.15) and InternVL2.5-78B (41.95 for 78B vs. 39.30 for 8B). Process-oriented supervision effectively offsets parameter scale limitations in visual reasoning tasks.

Per-dimension ablation.

Figure 8 details how each training variant affects the eight evaluation dimensions. The teaching trajectories approach yields the most pronounced gains in Canvas Collaboration dimensions, improving Spatio-Temporal Coherence by +20.3 points and Constructive Integrity by +14.1 points over the base model. "Text only" training provides minimal benefit in these specific areas, confirming that textual CoT supervision cannot teach spatial construction. "Image only" training provides negligible improvement across all dimensions and causes some regressions (HE drops from 44.31 to 43.90). TAV remains the only dimension where all variants perform comparably, as HTML/SVG syntax is relatively straightforward to learn from any training signal.

5.4 Data Efficiency Analysis

We evaluate whether the observed gains depend on large-scale data or benefit primarily from our curation strategy. Figure 9 presents performance as a function of training set size across eight checkpoints from 500 to 4,000 instances.

The initial gains are steep, as the 8B model achieves a CBS of 30.2 (an 11.7% relative improvement) with only 500 teaching trajectories. This confirms that a small

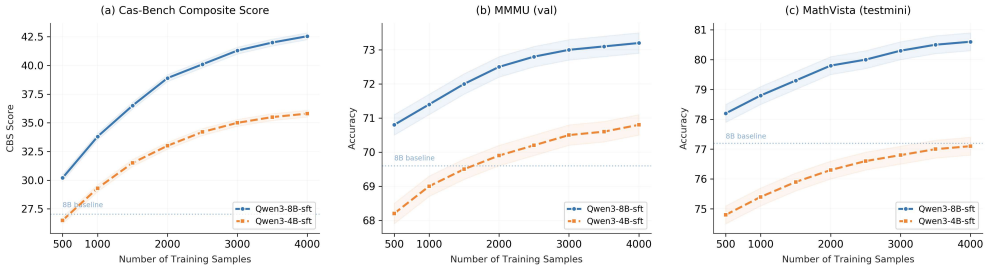
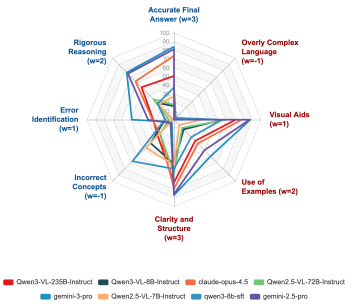


Fig. 9: Performance scaling with training data volume. Both the 4B and 8B student models show rapid initial gains that plateau around 3,000–3,500 samples, suggesting that 4,000 high-quality teaching trajectories provide near-saturating supervision. Dashed horizontal lines indicate base model performance (no fine-tuning).

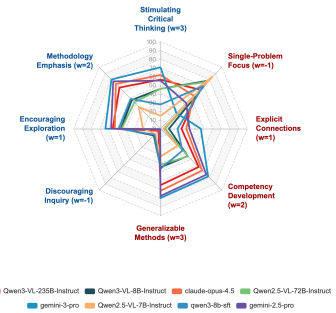
amount of aligned process data provides meaningful supervision. Diminishing returns appear around 3,000 samples, and the final 1,000 samples contribute only 1.25 CBS points for the 8B model. This scaling behavior remains consistent across CBS, MMMU, and MathVista for both model scales, indicating a fundamental property of the distillation framework rather than a task-specific artifact. Consequently, the quality of alignment, ensured by dynamic-rubric filtering, outweighs raw data volume.

5.5 Fine-Grained Analysis: Mathematics Domain

To investigate behavioral patterns, we conduct a fine-grained analysis using mathematics-specific rubric criteria (Figures 10 and 11). Mathematics serves as an ideal stress-test because it demands abstract symbolic manipulation alongside precise spatial-temporal alignment.

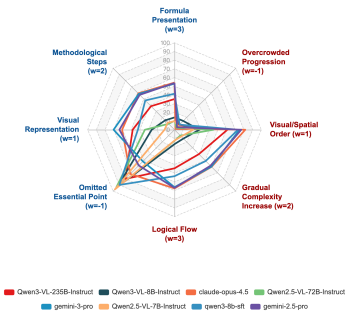


(a) Logical Rigor and Instructional Clarity

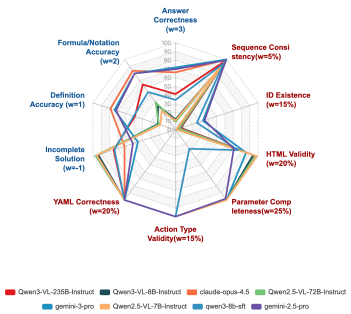


(b) Heuristic Engagement and Strategic Generalizability

Fig. 10: Fine-grained radar charts of Pedagogical Discourse in the mathematics domain.



(a) Constructive Integrity and Spatio-Temporal Coherence



(b) Semantic Fidelity and Technical Action Validation

Fig. 11: Fine-grained radar charts of Canvas Collaboration in the mathematics domain.

High-order logical rigor remains a proprietary advantage.

Gemini-3-Pro and Claude-Opus-4.5 dominate foundational metrics like Accurate Final Answer (84.1 and 73.6) and Rigorous Reasoning (76.8 and 62.5). Gemini-3-Pro is the only model demonstrating viable Error Identification capability (48.4, while others fall below 30.0). Open-source models lack the self-correcting deductive mechanisms needed to identify logical fallacies during the teaching process.

The spatial-symbolic bottleneck is the primary vulnerability.

The CC channel exposes the most significant weaknesses. Proprietary models maintain stable Formula Presentation (around 53–54) and Formula/Notation Accuracy (74–79), but Gemini-3-Pro is the only model to exceed 50 in comprehensive Visual Representation (70.2). Models broadly struggle to accurately map complex mathematical notations onto precise 2D canvas coordinates.

Syntax mastery does not imply semantic coherence.

A divergence emerges in Technical Action Validation, where almost all models perform well (Qwen3-VL-235B achieves 95.4 in HTML Validity). However, syntactic correctness does not guarantee semantic coherence. Claude-Opus-4.5 surpasses Gemini-3-Pro in Logical Flow (81.6 vs. 75.1), demonstrating that the core challenge lies in temporally synchronizing code with the evolving logical narrative rather than generating valid syntax. This finding supports our emphasis on temporal alignment during distillation.

5.6 Cross-Disciplinary Analysis

Figure 12 presents a cross-domain performance heatmap that captures the varying pedagogical requirements across disciplines. All models exhibit a performance gradient from Mathematics (most demanding) through Natural Sciences and Engineering to Humanities (most forgiving). For proprietary models, this gradient is moderate; Gemini-3-Pro drops from 72.5 in Mathematics to 66.5 in Humanities. For open-source

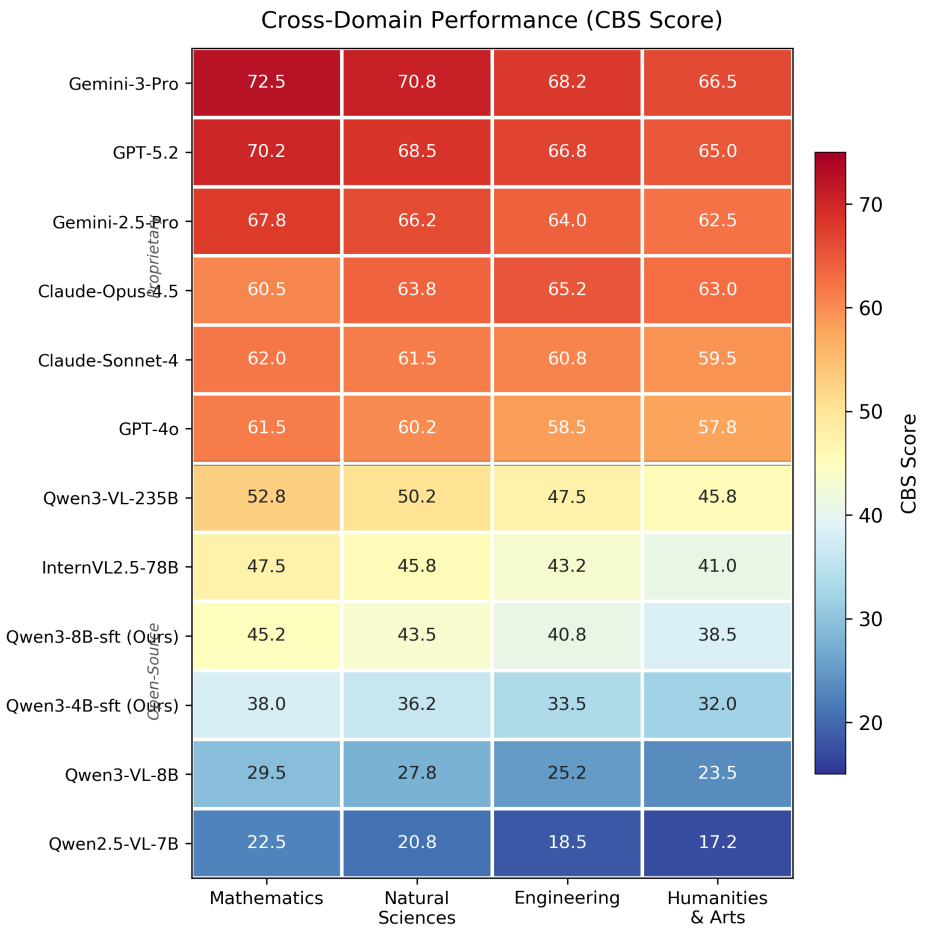


Fig. 12: Cross-domain CBS performance heatmap. The black horizontal line separates proprietary (above) from open-source (below) models. All models show a consistent performance gradient from Mathematics (hardest) to Humanities (most forgiving), with the gap widening for weaker models.

models, the gradient is steeper, with Qwen3-VL-8B dropping from 29.5 to 23.5 (a 20.3% relative decline).

In STEM tasks, the evaluation heavily penalizes logical gaps. Mathematical teaching relies on exact spatial mapping and rigid deductive chains, requiring stringent Logical Rigor and precise Constructive Integrity. In Humanities tasks, the pedagogical focus shifts to conceptual exploration and relational mapping. The canvas is typically used for mind maps, timelines, or thematic comparisons rather than equations. HTML/SVG constraints are more forgiving in these domains, shifting the emphasis from mathematical accuracy to semantic organization.

Our fine-tuned models maintain relatively consistent cross-domain performance. Qwen3-8B-sft achieves 45.2 in Mathematics and 38.5 in Humanities. This narrower gap compared to similarly sized base models suggests that the teaching trajectories, which span all four domains, provide balanced cross-disciplinary supervision.

6 Conclusion

In this paper, we have presented a dual-stream pedagogical distillation framework for interleaved visual-text reasoning, grounded in the "Explain-on-Blackboard" paradigm. Our core methodological contribution lies in the principled enforcement of temporal and semantic alignment between textual reasoning and visual construction during knowledge distillation from frontier teacher models. By drawing on cognitive theories of multimedia learning, dual coding, and social constructivism, we establish that pedagogical explanation provides the ideal cognitive scaffold for generating high-quality, aligned training data.

Our empirical evaluation reveals a significant representation tax in current models, quantifying the cognitive overhead of translating abstract logic into structured spatial representations. The proposed dynamic-rubric quality assurance mechanism provides a scalable solution for curating process-oriented training data, achieving over 94% alignment with human expert judgments. Through ablation studies, we demonstrate that the alignment between textual and visual streams, rather than either modality in isolation, is the critical factor for effective interleaved reasoning.

The practical impact of our approach is substantial: a lightweight 8B-parameter model fine-tuned on 4,000 distilled trajectories achieves a 69.76% relative improvement on Cas-Bench and consistent gains across four established multimodal benchmarks, validating the effectiveness of process-oriented supervision. These results suggest that structured knowledge distillation can effectively compensate for parameter scale, providing a scalable pathway for endowing compact models with interpretable, human-like visual reasoning capabilities.

Future work may extend this framework to interactive teaching environments, explore reinforcement learning from process feedback, and develop domain-adaptive distillation strategies that account for the epistemological differences between STEM and humanities tasks.

Declarations

Funding This research is supported in part by the National Natural Science Foundation of China under Grant No. 12071160, and the Natural Science Foundation of Guangdong Province under Grant No. 2020A1515010566.
Conflict of interest/Competing interests Not applicable
Ethics approval and consent to participate Not applicable
Consent for publication Not applicable
Data availability The datasets generated and analyzed during the current study, specifically the 4,000 distilled teaching trajectories and the Cas-Bench evaluation protocol, will be made publicly available upon the acceptance and publication of this manuscript. The foundational problems comprising Cas-Bench are derived from the

publicly available CoMT, MathVision, and MMMU-Pro benchmarks. Prior to publication, the datasets are available from the corresponding author upon reasonable request.

Materials availability Not applicable

Code availability Prior to publication, the source code is available from the corresponding author upon reasonable request.

Author contribution Linzhuang Sun: Led the algorithmic and system design and contributed to manuscript writing. Sizhe Qiu: Implemented experiments and contributed to manuscript writing. Jianwei Cao: Validated experiments and contributed to manuscript writing. Bihui Yu: Project administration and supervision. Zhaojin Fu: Manuscript revision and writing. Taishan Zeng: Writing – review & editing, and supervision.

Appendix A Prompts

The evaluation prompts and rubric generation prompts are shown in Figure A1, Figure A2, Figure A3, Figure A4, Figure A5, Figure A6, Figure A7, and Figure A8.

Evaluator Prompt (PD)

SYSTEM

You are a rigorous and critical evaluator. Given a question, a model-generated teaching explanation, the correct standard answer, the correct solution, and a detailed list of evaluation rubrics, you must evaluate the explanation against EACH rubric individually with strict adherence to the scoring guidelines below.

[Critical Scoring Guidelines & Anti-Inflation Measures]

1. Granular Scoring Scale with Forced Differentiation:

100 (Perfect): Virtually never assigned.The explanation must be flawless and represent a pinnacle of pedagogical excellence that cannot be realistically improved upon in any aspect related to the rubric. If you consider 100, you must justify why it is impossible to make it even slightly better for a learner

90 (Excellent): Reserved for explanations that are outstanding but have a “trivial, non-substantive flaw” (e.g., a single slightly awkward phrase that does not impede understanding). This is still a high bar.

80 (Good): The explanation is solid and effective in fulfilling the rubric's intent. However, it has at least one identifiable, specific minor shortcoming (e.g., a step explained slightly too quickly, a useful analogy not fully developed, a non-critical detail omitted). An "80" must have a concrete deficiency mentioned in the reasoning.

70 (Satisfactory+): Meets the core requirement of the rubric competently but is unremarkable. It lacks the depth, insight, or polished pedagogical flow that would make it "Good." It's correct but basic.

60 (Satisfactory-): Barely meets the basic requirements. The fulfillment of the rubric is partial, unclear, or superficially addressed. May contain minor logical jumps or lack emphasis on important aspects.

50 (Weak): Addresses the rubric's theme but with significant gaps in clarity, completeness, or accuracy. The explanation is muddled or incomplete on key points.

40 or below (Poor/Unacceptable): Fails to meet the basic intent of the rubric. May be missing, incorrect, or misleading.

2. Mandatory "Ceiling Rules" to Prevent Inflation:

"Mentioned vs. Mastered" Rule: If the explanation only mentions or states the concept required by the rubric but does not explain, elaborate, or connect it pedagogically, its score MUST NOT exceed 70. Presence alone is not excellence.

"Factual Error" Rule: If the explanation contains any factual inaccuracy, logical contradiction, or misleading statement relevant to the rubric, its score MUST NOT exceed 40, regardless of other qualities.

"Default to Middle" Bias Correction: You must actively search for shortcomings. An explanation without obvious errors should not default to a high score (90+). Ask: "Could this be clearer? More insightful? Better structured for a student?" If yes, cap the score at 80 or lower.

3. Strict Application of Calibration Rule:

The rule "If a criterion is merely present but not explicitly emphasized or pedagogically strong, assign a score between 50 and 80" is MANDATORY. Do not reward box-ticking. Reserve scores above 80 for demonstrably strong teaching moves.

Fig. A1: Prompt for Pedagogical Discourse Evaluation (1/2).

Evaluator Prompt (PD)

USER

Given the following question, explanation, correct standard answer, correct solution, and rubrics, evaluate each rubric individually with strict adherence to the Anti-Inflation Measures above and calculate the final weighted score.

<question> {question} </question>
<explanation> {explanation} </explanation>
<standard_ans>
{standard_ans}
</standard_ans>
<solution> {solution} </solution>
<rubrics> {rubric_list_string} </rubrics>

Evaluation Instructions:

For each rubric item in the provided list, output:
rubric_title: [The title of the rubric]
satisfaction_score: [A number from 0 to 100, determined STRICTLY by the Scoring Guidelines]
reasoning: [A concrete 1-2 sentence explanation. For scores >= 80, you MUST state the minor flaw. For scores <= 70, you MUST state the missing depth or specific gap.]

After evaluating all rubrics, calculate the final weighted score using this formula:
Final Score = Σ(weight × satisfaction_score) / Σ(positive_weights_only)
Only sum the weights that are positive (Essential, Important, Optional rubrics).
Pitfall rubrics (with negative weights) only contribute to the numerator as penalties.

Important note about Pitfall rubrics:
For rubrics with negative weights (Pitfall categories):
- A satisfaction_score of 0 means the pitfall is completely avoided (no penalty).
- A satisfaction_score of 100 means the pitfall is fully triggered (full penalty).
- Intermediate scores represent partial triggering of the pitfall.

Overall Score Expectation Context:
- A final score above 85 should correspond to an overall excellent explanation.
- A final score between 70 and 84 should correspond to a good to very good explanation.
- A final score between 60 and 69 should correspond to an adequate but flawed explanation.
- A final score below 60 indicates a substantially poor explanation.
Use this context to sanity-check your rubric-level scores.

Output format must be a valid JSON object with:
rubric_evaluations: A list of objects, each containing `rubric_title`, `satisfaction_score`, and `reasoning`.
final_score: The calculated final weighted score (rounded to 3 decimal places).

Your JSON evaluation:

Fig. A2: Prompt for Pedagogical Discourse Evaluation (2/2).

Evaluator Prompt (CC)

SYSTEM

You are a rigorous and critical evaluator. Given a question, a model-generated teaching explanation, the correct standard answer, the correct solution, and a detailed list of evaluation rubrics, you must evaluate the explanation against EACH rubric individually with strict adherence to the scoring guidelines below.

[Critical Scoring Guidelines & Anti-Inflation Measures]

1. Granular Scoring Scale with Forced Differentiation:

100 (Perfect): Virtually never assigned.The explanation must be flawless and represent a pinnacle of pedagogical excellence that cannot be realistically improved upon in any aspect related to the rubric. If you consider 100, you must justify why it is impossible to make it even slightly better for a learner

90 (Excellent): Reserved for explanations that are outstanding but have a “trivial, non-substantive flaw” (e.g., a single slightly awkward phrase that does not impede understanding). This is still a high bar.

80 (Good): The explanation is solid and effective in fulfilling the rubric's intent. However, it has at least one identifiable, specific minor shortcoming (e.g., a step explained slightly too quickly, a useful analogy not fully developed, a non-critical detail omitted). An "80" must have a concrete deficiency mentioned in the reasoning.

70 (Satisfactory+): Meets the core requirement of the rubric competently but is unremarkable. It lacks the depth, insight, or polished pedagogical flow that would make it "Good." It's correct but basic.

60 (Satisfactory-): Barely meets the basic requirements. The fulfillment of the rubric is partial, unclear, or superficially addressed. May contain minor logical jumps or lack emphasis on important aspects.

50 (Weak): Addresses the rubric's theme but with significant gaps in clarity, completeness, or accuracy. The explanation is muddled or incomplete on key points.

40 or below (Poor/Unacceptable): Fails to meet the basic intent of the rubric. May be missing, incorrect, or misleading.

2. Mandatory "Ceiling Rules" to Prevent Inflation:

"Mentioned vs. Mastered" Rule: If the explanation only mentions or states the concept required by the rubric but does not explain, elaborate, or connect it pedagogically, its score MUST NOT exceed 70. Presence alone is not excellence.

"Factual Error" Rule: If the explanation contains any factual inaccuracy, logical contradiction, or misleading statement relevant to the rubric, its score MUST NOT exceed 40, regardless of other qualities.

"Default to Middle" Bias Correction: You must actively search for shortcomings. An explanation without obvious errors should not default to a high score (90+). Ask: "Could this be clearer? More insightful? Better structured for a student?" If yes, cap the score at 80 or lower.

3. Strict Application of Calibration Rule:

The rule "If a criterion is merely present but not explicitly emphasized or pedagogically strong, assign a score between 50 and 80" is MANDATORY. Do not reward box-ticking. Reserve scores above 80 for demonstrably strong teaching moves.

Fig. A3: Prompt for Canvas Collaboration Evaluation (1/2).

Evaluator Prompt (CC)

USER

Given the following question, correct standard answer, correct solution, and rubrics, evaluate ONLY THE CONTENT IN THE PROVIDED IMAGE(S) against each rubric individually with strict adherence to the Anti-Inflation Measures above and calculate the final weighted score.

<question> {question} </question>
<standard_ans>
{standard_ans}
</standard_ans>
<solution> {solution} </solution>
<rubrics> {rubric_list_string} </rubrics>

Evaluation Instructions:

For each rubric item in the provided list, output:
rubric_title: [The title of the rubric]
satisfaction_score: [A number from 0 to 100, determined STRICTLY by the Scoring Guidelines based SOLELY on the image(s)]
reasoning: [A concrete 1-2 sentence explanation based only on the image(s). For scores >= 80, you MUST state the minor visual flaw. For scores <= 70, you MUST state the missing visual depth or specific gap.]

After evaluating all rubrics, calculate the final weighted score using this formula:
Final Score = Σ(weight × satisfaction_score) / Σ(positive_weights_only)
Only sum the weights that are positive (Essential, Important, Optional rubrics).
Pitfall rubrics (with negative weights) only contribute to the numerator as penalties.

Important note about Pitfall rubrics:
For rubrics with negative weights (Pitfall categories):
- A satisfaction_score of 0 means the pitfall is completely avoided (no penalty) in the visual content.
- A satisfaction_score of 100 means the pitfall is fully triggered (full penalty) by the visual content.
- Intermediate scores represent partial triggering of the pitfall by the visual content.

Overall Score Expectation Context:
- A final score above 85 should correspond to overall excellent visual explanations.
- A final score between 70 and 84 should correspond to good to very good visual explanations.
- A final score between 60 and 69 should correspond to adequate but visually flawed explanations.
- A final score below 60 indicates a substantially poor visual explanation.
Use this context to sanity-check your rubric-level scores.

Critical Focus Notice:Your evaluation MUST BE SOLELY AND EXPLICITLY based on the content of the image(s). Ignore the quality of the accompanying text explanation. If a concept is described in text but not effectively visualized, the visual content should be penalized accordingly.

Output format must be a valid JSON object with:
rubric_evaluations: A list of objects, each containing `rubric_title`, `satisfaction_score`, and `reasoning`.
final_score: The calculated final weighted score (rounded to 3 decimal places).

Your JSON evaluation:

Fig. A4: Prompt for Canvas Collaboration Evaluation (2/2).

Rubric generation Prompt (PD)

You are an expert rubric designer for evaluating teaching explanations. Your task is to create a single, unified set of evaluation criteria (a "rubric") for assessing the quality of classroom-style teaching explanations provided by teachers or AI tutors.

The rubric should be general enough to apply to all mathematics teaching explanations, yet specific enough to be useful for detailed evaluation. It should encompass common patterns and best practices found in effective teaching explanations.

Input:
- 30 sample mathematics teaching explanations

Your task is to analyze these sample explanations and generate a single, comprehensive rubric that can be used to evaluate any mathematics teaching explanation.

The rubric must be structured around the following four core dimensions, each containing several specific criteria:

1. Logical Rigor
   Evaluates whether the answer is accurate, the reasoning process is rigorous and aligns with academic standards. If the problem has multiple solutions, it assesses whether all are provided. Evaluation should be based on the standard answer.

2. Instructional Clarity
   Evaluates whether the explanation is clear and easy to comprehend, including language expression, logical structure, step-by-step breakdown, and whether appropriate analogies or examples are used to aid understanding.

3. Heuristic Engagement
   Evaluates whether the explanation stimulates student thinking, including posing guiding questions, offering unique perspectives, emphasizing methodology, and using encouraging language to motivate student exploration.

4. Strategic Generalizability
   Evaluates whether the expanation aims to cultivate students' long-term, sustainable learning abilities, including guiding them to establish generalizable methods, expanding their thinking boundaries, and focusing on competency development rather than solving a single problem.

Overall Requirements:
A teacher's ability to provide mathematical teaching explanations will be evaluated across these four dimensions. You need to generate exactly 4 specific rubric items for each dimension, making a total of 16 items.
For each dimension, the 4 items must follow this specific weight distribution pattern:
1. One item with weight = 3 (Essential)
2. One item with weight = 2 (Important)
3. One item with weight = 1 (Optional)
4. One item with weight = -1 (Pitfall)
Each rubric item should be as independent as possible from the others, assessing a distinct aspect of the explanation.

Format Requirements for Each Rubric Item:
Each item must contain exactly three key components:
1. Title (2–4 words)
2. Description: One sentence beginning with its category prefix, explicitly stating what to look for.
3. Weight: Must be exactly one of: 3, 2, 1, or -1

Fig. A5: Prompt for Pedagogical Discourse Rubric Generation (1/2).

Rubric generation Prompt (PD)

Guidance on Item Categories:
- Weight 3 (Essential): Core instructional elements; the explanation fails as a teaching attempt without them (e.g., incorrect concepts, missing key reasoning steps).
- Weight 2 (Important): Elements that significantly enhance student understanding and instructional quality.
- Weight 1 (Optional): Helpful teaching enhancements (e.g., examples, analogies, intuitive explanations) that are beneficial but not critical.
- Weight -1 (Pitfall): Common teaching failures, such as skipping reasoning steps, presenting conclusions without explanation, or using language inappropriate for the intended student level.

Format and Structure Specifications:
- Overall Structure: Your final output must clearly be divided into four sections (Correctness, Understandability, Insightfulness, Rooting). Each section should contain an array of exactly 4 rubric items, with weights distributed as specified above (3, 2, 1, -1).
- Item Distribution: Allocate items appropriately within each dimension based on their characteristics. For example, the "Correctness" dimension's Essential item (weight 3) might focus on fundamental accuracy, while the "Insightfulness" dimension's Essential item (weight 3) might focus on stimulating critical thinking.
- Content Requirements: Ensure each dimension has a balanced representation of different aspects. Avoid including style-only preferences unless they demonstrably affect student understanding.
- Self-Containment & Independence: Each rubric item must be fully self-contained and assess a distinct, non-overlapping aspect of the explanation so that non-expert readers (e.g., students or annotators) can understand and use it without needing to consult any external information or context.

Output Format:
Please provide a clear JSON object as the output. This object must contain four keys corresponding to the four dimensions: `correctness`, `understandability`, `insightfulness`, `rooting`. The value for each key is a JSON array containing exactly 4 rubric item objects for that dimension, with weights 3, 2, 1, -1 respectively.
Each item object must contain exactly three keys: `title`, `description`, and `weight`.
Each description must begin with its corresponding category prefix based on weight:
- Weight 3: `[Essential]`
- Weight 2: `[Important]`
- Weight 1: `[Optional]`
- Weight -1: `[Pitfall]`
No extra keys are allowed.
Do not copy large blocks of text from the sample explanations into the rubric descriptions. Use the sample explanations only to identify common patterns and best practices.

Now, based on the given 30 sample explanations, generate a unified, dimensionally structured rubric according to the specified structure, dimensions, and requirements.

Fig. A6: Prompt for Pedagogical Discourse Rubric Generation (2/2).

Rubric generation Prompt (CC)

You are an expert rubric designer for evaluating the teaching effectiveness of classroom board content (e.g., blackboard, whiteboard, or structured visual board representations). Your task is to create a single, unified set of evaluation criteria (a "rubric") for assessing how effectively board content supports student learning in mathematics.

The rubric should be general enough to apply to all mathematics board content, yet specific enough to be useful for detailed evaluation. It should encompass common patterns and best practices found in effective board usage.

Input:
- 30 sample mathematics board contents

Your task is to analyze these sample board contents and generate a single, comprehensive rubric that can be used to evaluate any mathematics board content.

The rubric must be structured around the following three core dimensions, each containing several specific criteria:

1. Constructive Integrity
Evaluates the breadth and depth of subject matter presented on the board. This includes whether all key concepts, definitions, formulas, and methodological steps relevant to the problem or lesson are clearly and adequately addressed.

2. Spatio-Temporal Coherence
Evaluates how logically the content unfolds on the board to support student understanding. This includes:
  a. From Simple to Complex: Whether the presentation starts with basic concepts and gradually moves to more difficult ones.
  b. Conceptual Coherence: Whether the logical connections between consecutive steps and concepts are made clear.
  c. Gradual Increase in Complexity: Whether the difficulty and abstraction level increase smoothly and purposefully across the board.

3. Semantic Fidelity
Evaluates the correctness and precision of all information presented on the board. This includes:
  a. Formula/Symbol Accuracy: Correctness of all mathematical formulas, symbols, and notation.
  b. Definition Statement Accuracy: Precision and correctness of all definitions, theorems, and rules stated.
  c. Data/Numerical Correctness: Accuracy of all numbers, calculations, and data points.
  d. Answer Completeness & Correctness: For problems requiring answers, the final result(s) must be correct and complete (covering all possible solutions if applicable).

Overall Requirements:
You need to generate exactly 4 specific rubric items for each of the three dimensions listed above, making a total of 12 items.
For each dimension, the 4 items must follow this specific weight distribution pattern:
1. One item with weight = 3 (Essential)
2. One item with weight = 2 (Important)
3. One item with weight = 1 (Optional)
4. One item with weight = -1 (Pitfall)
Each rubric item should be as independent as possible from the others, assessing a distinct aspect of the board content.

Format Requirements for Each Rubric Item:
Each item must contain exactly three key components:
1. Title: (2–4 words)
2. Description: One sentence beginning with its category prefix, explicitly stating what to look for in the board content.
3. Weight: Must be exactly one of: 3, 2, 1, or -1

Fig. A7: Prompt for Canvas Collaboration Rubric Generation (1/2).

Rubric generation Prompt (CC)

Guidance on Item Categories:
- Weight 3 (Essential): Fundamental board-level requirements; the board fails as a teaching aid without them (e.g., missing key intermediate steps, structurally illegible, or containing incorrect information).
- Weight 2 (Important): Board practices that significantly enhance instructional clarity, logical flow, and learning effectiveness.
- Weight 1 (Optional): Helpful visual or organizational enhancements that improve comprehension but are not strictly necessary (e.g., strategic spacing, highlighting, grouping).
- Weight -1 (Pitfall): Common board-related teaching failures, such as writing only the final answer, overcrowding content, presenting steps in an order inconsistent with the explanation, or introducing logical leaps without justification.

Format and Structure Specifications:
- Overall Structure: Your final output must clearly be divided into three sections (Knowledge Point Coverage, Instructional Progression Rationality, Factual Accuracy). Each section should contain a JSON array of exactly 4 rubric items, with weights distributed as specified above (3, 2, 1, -1).
- Item Distribution: Allocate the items appropriately based on the characteristics of each dimension. For example, the "Factual Accuracy" dimension's Essential item (weight 3) might focus on core accuracy, while "Instructional Progression Rationality" dimension's Essential item (weight 3) might focus on logical flow.
- Content Requirements: Ensure each dimension covers its specific sub-aspects appropriately across the different weight categories. For example, under "Factual Accuracy," include criteria for formulas, definitions, data, and final answers distributed across the 3, 2, 1, -1 weights.
- Self-Containment & Independence: Each rubric item must be fully self-contained and assess a distinct, non-overlapping aspect of the board content so that non-expert readers (e.g., students or annotators) can understand and use it without needing to consult any external information or context.

Output Format:
Please provide a clear JSON object as the output. This object must contain three keys corresponding to the three dimensions: `knowledge_point_coverage`, `instructional_progression_rationality`, `factual_accuracy`. The value for each key is a JSON array containing exactly 4 rubric item objects for that dimension, with weights 3, 2, 1, -1 respectively.
Each item object must contain exactly three keys: `title`, `description`, and `weight`.
Each description must begin with its corresponding category prefix based on weight:
- Weight 3: `[Essential]`
- Weight 2: `[Important]`
- Weight 1: `[Optional]`
- Weight -1: `[Pitfall]`
No extra keys are allowed.
Do not copy large blocks of text from the sample board contents into the rubric descriptions. Use the sample board contents only to identify common patterns and best practices.

Now, based on the given 30 sample board contents, generate a unified, dimensionally structured rubric according to the specified structure, dimensions, and requirements.

Fig. A8: Prompt for Canvas Collaboration Rubric Generation (2/2).

References

[1] Wang, Y., Chen, Q., Li, Z., Wang, S., Guo, S., Zhang, Z., Wei, Z.: Simple o3: Towards interleaved vision-language reasoning. arXiv preprint arXiv:2508.12109 (2025)

[2] Chen, X., Zhang, R., Jiang, D., Zhou, A., Yan, S., Lin, W., Li, H.: Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. arXiv preprint arXiv:2506.05331 (2025)

[3] Hu, Y., Shi, W., Fu, X., Roth, D., Ostendorf, M., Zettlemoyer, L., Smith, N.A., Krishna, R.: Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. Advances in Neural Information Processing Systems **37**, 139348–139379 (2024)

[4] Liu, C., Yang, Y., Fan, Y., Wei, Q., Liu, S., Wang, X.E.: Reasoning within the mind: Dynamic multimodal interleaving in latent space. arXiv preprint arXiv:2512.12623 (2025)

[5] Pham, T.-H., Ngo, C.: Multimodal chain of continuous thought for latent-space reasoning in vision-language models. arXiv preprint arXiv:2508.12587 (2025)

[6] Cheng, Z., Chen, Q., Zhang, J., Fei, H., Feng, X., Che, W., Li, M., Qin, L.: Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 23678–23686 (2025)

[7] Gu, J., Hao, Y., Wang, H.W., Li, L., Shieh, M.Q., Choi, Y., Krishna, R., Cheng, Y.: Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. In: The Fourteenth International Conference on Learning Representations (2025)

[8] Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., *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)

[9] Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)

[10] Li, J., Lu, W., Fei, H., Luo, M., Dai, M., Xia, M., Jin, Y., Gan, Z., Qi, D., Fu, C., et al.: A survey on benchmarks of multimodal large language models. arXiv preprint arXiv:2408.08632 (2024)

[11] Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench: Benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13299–13308 (2024)

[12] Chi, M.T., De Leeuw, N., Chiu, M.-H., LaVancher, C.: Eliciting self-explanations improves understanding. Cognitive science **18**(3), 439–477 (1994)

[13] Mayer, R.E.: The Cambridge Handbook of Multimedia Learning. Cambridge university press, ??? (2005)

[14] Paivio, A.: Dual coding theory: Retrospect and current status. Canadian Journal of Psychology/Revue canadienne de psychologie **45**(3), 255 (1991)

[15] Vygotsky, L.S., Cole, M.: Mind in Society: Development of Higher Psychological Processes. Harvard university press, ??? (1978)

[16] Wang, K., Pan, J., Shi, W., Lu, Z., Ren, H., Zhou, A., Zhan, M., Li, H.: Measuring multimodal mathematical reasoning with math-vision dataset. Advances in Neural Information Processing Systems **37**, 95095–95169 (2024)

[17] Yue, X., Zheng, T., Ni, Y., Wang, Y., Zhang, K., Tong, S., Sun, Y., Yu, B., Zhang, G., Sun, H., *et al.*: Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 15134–15186 (2025)

[18] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., *et al.*: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)

[19] Wang, Y., Wu, S., Zhang, Y., Yan, S., Liu, Z., Luo, J., Fei, H.: Multimodal chain-of-thought reasoning: A comprehensive survey. arXiv preprint arXiv:2503.12605 (2025)

[20] Hao, S., Gu, Y., Ma, H., Hong, J., Wang, Z., Wang, D., Hu, Z.: Reasoning with language model is planning with world model. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 8154–8173 (2023)

[21] Ma, W., Chen, H., Zhang, G., Chou, Y.-C., Chen, J., Melo, C., Yuille, A.: 3dsr-bench: A comprehensive 3d spatial reasoning benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6924–6934 (2025)

[22] Wu, J., Guan, J., Feng, K., Liu, Q., Wu, S., Wang, L., Wu, W., Tan, T.: Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. arXiv preprint arXiv:2506.09965 (2025)

[23] Chen, S., Zhu, T., Zhou, R., Zhang, J., Gao, S., Niebles, J.C., Geva, M., He, J., Wu, J., Li, M.: Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. arXiv preprint arXiv:2503.01773 (2025)

[24] Weng, S., Wang, Z., Zhou, Y., Lu, R., Liu, T., Teng, Z., Liu, X., Liu, H.: Geosketch: A neural-symbolic approach to geometric multimodal reasoning with auxiliary line construction and affine transformation. arXiv preprint arXiv:2509.22460 (2025)

[25] Sun, L., Zhu, Y., Liu, R., Liang, H., Sun, Z., Jia, C., He, H., Wu, Y., Li, S., Wei, J., et al.: Canvas-of-thought: Grounding reasoning via mutable structured states. arXiv preprint arXiv:2602.10494 (2026)

[26] Zha, D., Bhat, Z.P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., Hu, X.: Data-centric artificial intelligence: A survey. ACM Computing Surveys **57**(5), 1–42 (2025)

[27] Yang, C., Zhu, Y., Lu, W., Wang, Y., Chen, Q., Gao, C., Yan, B., Chen, Y.: Survey on knowledge distillation for large language models: methods, evaluation, and application. ACM Transactions on Intelligent Systems and Technology **16**(6), 1–27 (2025)

[28] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)

[29] Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26296–26306 (2024)

[30] Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. In: European Conference on Computer Vision, pp. 370–387 (2024). Springer

[31] Sun, L., Liang, H., Wei, J., Yu, B., Li, T., Yang, F., Zhou, Z., Zhang, W.: Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 14100–14115 (2025)

[32] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), 403 (2024)

[33] Wu, J., Gan, W., Chen, Z., Wan, S., Yu, P.S.: Multimodal large language models: A survey. In: 2023 IEEE International Conference on Big Data (BigData), pp. 2247–2256 (2023). IEEE

[34] Lu, P., Gong, R., Jiang, S., Qiu, L., Huang, S., Liang, X., Zhu, S.-C.: Intergps: Interpretable geometry problem solving with formal language and symbolic

reasoning. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 6774–6786 (2021)

[35] Kahou, S.E., Michalski, V., Atkinson, A., Kádár, Á., Trischler, A., Bengio, Y.: Figureqa: An annotated figure dataset for visual reasoning. arXiv preprint arXiv:1710.07300 (2017)

[36] Li, Z., Wang, X., Stengel-Eskin, E., Kortylewski, A., Ma, W., Van Durme, B., Yuille, A.L.: Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14963–14973 (2023)

[37] Lu, P., Qiu, L., Chang, K.-W., Wu, Y.N., Zhu, S.-C., Rajpurohit, T., Clark, P., Kalyan, A.: Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. arXiv preprint arXiv:2209.14610 (2022)