Unveiling Fine-Grained Visual Traces: Evaluating Multimodal Interleaved Reasoning Chains in Multimodal STEM Tasks

Jing Jin[1,2*] **Hao Liu**[1*] **Yan Bai**[2*] **Yihang Lou**[4] **Zhenke Wang**[3] **Tianrun Yuan**[3]
Juntong Chen[3] **Yongkang Zhu**[3] **Fanhu Zeng**[5] **Xuanyu Zhu**[2,4] **Tao Feng**[1†] **Yige Xu**[6†]

[1]Tsinghua University [2]Meituan Inc [3]Central South University [4]Peking University
[5]University of Chinese Academy of Sciences [6]Nanyang Technological University
jingjin0007@gmail.com liuhaokix@csu.edu.cn yige002@ntu.edu.sg

Abstract

Multimodal large language models (MLLMs) have shown promising reasoning abilities, yet evaluating their performance in specialized domains remains challenging. STEM reasoning is a particularly valuable testbed because it provides highly verifiable feedback, but existing benchmarks often permit unimodal shortcuts due to modality redundancy and focus mainly on final-answer accuracy, overlooking the reasoning process itself. To address this challenge, we introduce **STEPSTEM**: a graduate-level benchmark of 283 problems across mathematics, physics, chemistry, biology, and engineering for fine-grained evaluation of cross-modal reasoning in MLLMs. **STEPSTEM** is constructed through a rigorous curation pipeline that enforces strict complementarity between textual and visual inputs. We further propose a general step-level evaluation framework for both text-only chain-of-thought and interleaved image-text reasoning, using dynamic programming to align predicted reasoning steps with multiple reference solutions. Experiments across a wide range of models show that current MLLMs still rely heavily on textual reasoning, with even Gemini 3.1 Pro and Claude Opus 4.6 achieving only 38.29% accuracy. These results highlight substantial headroom for genuine cross-modal STEM reasoning and position **STEPSTEM** as a benchmark for fine-grained evaluation of multimodal reasoning. The benchmark and evaluation code are available at `https://github.com/lll-hhh/STEPSTEM`.

1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) [43, 35, 8, 5, 7, 40] have greatly expanded the scope of multimedia understanding and generation. Beyond processing complex visual and textual inputs, MLLMs can now produce intermediate reasoning traces in textual, visual, or interleaved image-text forms, and use them to support final predictions [39, 20, 6]. This capability makes MLLMs a promising foundation for complex multimodal reasoning. Among such tasks, STEM reasoning (Science, Technology, Engineering, and Mathematical) is particularly important because final answers are often objectively verifiable, providing a natural testbed for analyzing both reasoning outcomes and reasoning processes.

Despite recent progress, multimodal STEM reasoning remains challenging. First, existing MLLMs often suffer from *modality collapse*: they over-rely on textual cues while under-utilizing visual evi-

* Equal contribution.
† Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

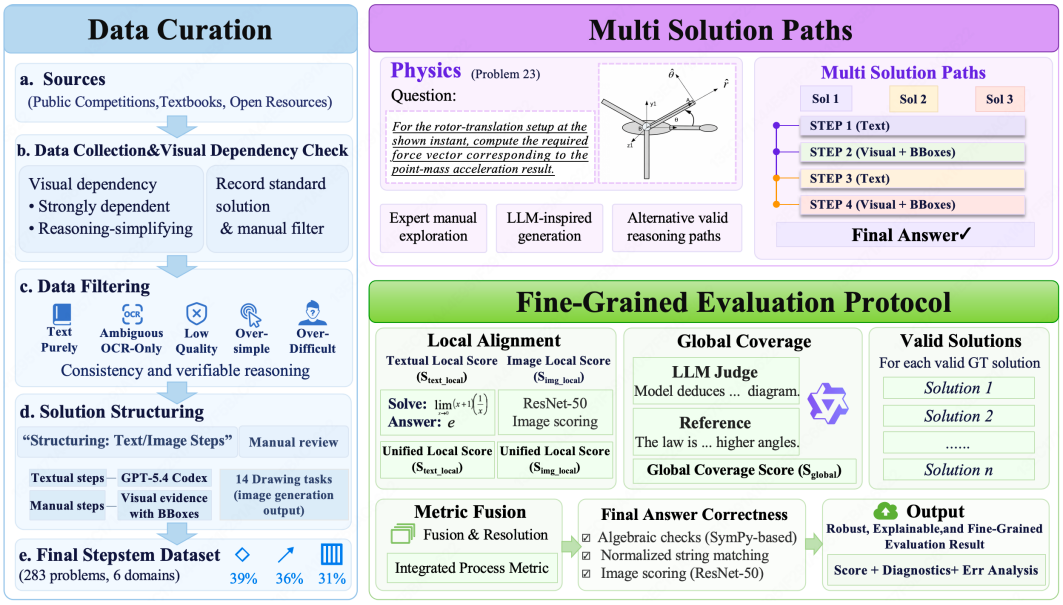

Figure 1: Overview of the StepSTEM construction and evaluation workflow, including the data curation pipeline, the multi-solution construction pipeline, and the fine-grained evaluation protocol.

dence, resulting in weak visual grounding and degraded performance on tasks that require genuine visual reasoning [55, 47, 49]. This issue is especially pronounced in STEM benchmarks, where detailed textual descriptions may allow models to bypass visual information. Second, current models still struggle to understand and generate structured STEM visuals. Scientific figures, diagrams, and plots are underrepresented in existing training data, and mainstream image-generation pipelines are largely optimized for naturalistic images rather than layouts with precise symbols, labels, and logical constraints [23, 45]. Third, current evaluation and post-training practices mainly emphasize final-answer correctness, while providing limited insight into the quality and faithfulness of intermediate reasoning. Prior work has shown that chain-of-thought rationales may not faithfully reflect the underlying reasoning process, and that step-wise supervision can better support reliable reasoning than outcome-only signals [41, 27, 52]. As a result, it remains difficult to determine whether a model truly performs grounded cross-modal reasoning or merely exploits superficial shortcuts.

To address these limitations, we introduce **StepSTEM**, a multimodal STEM benchmark with fine-grained reasoning-step evaluation. StepSTEM contains 283 graduate-level problems spanning engineering, physics, mathematics, chemistry, and biology. Each problem is carefully curated to require both textual information and visual evidence, thereby discouraging unimodal shortcuts and encouraging genuine cross-modal reasoning. As shown in Figure 1, our data construction process enforces strict text-vision complementarity: solutions cannot be recovered from text alone or vision alone. Moreover, each problem is annotated with one or more ground-truth solution trajectories containing both textual and visual intermediate steps.

Meanwhile, we propose a *step-level* evaluation framework for multimodal reasoning. Rather than evaluating only final-answer accuracy, our framework measures the similarity between predicted reasoning steps and annotated ground-truth steps, then aligns predicted trajectories with one or more reference trajectories using dynamic programming. The resulting score assesses not only whether a model obtains the correct answer, but also whether it follows a plausible, grounded, and interpretable multimodal reasoning path.

We evaluate a broad range of representative MLLMs on StepSTEM. Experimental results show that existing models still struggle with graduate-level multimodal STEM reasoning, revealing substantial room for improvement. In addition, our step-level scores provide a more fine-grained view of model behavior than answer-only evaluation, exposing differences in reasoning quality, modality usage, and grounding ability that are obscured by final accuracy alone.

In summary, our contributions are three-fold: **(1) A multimodal STEM reasoning benchmark requiring cross-modal reasoning:** We introduce **STEPSTEM**, a benchmark of 283 carefully curated STEM problems that enforce strict text-vision complementarity to reduce unimodal shortcuts. **(2) A fine-grained framework for intermediate reasoning evaluation:** We propose a **step-level evaluation framework** that systematically assesses intermediate multimodal reasoning steps beyond final-answer accuracy; and **(3) A comprehensive empirical analysis:** We evaluate representative MLLMs on **STEPSTEM** and analyze their capabilities and limitations in cross-modal STEM reasoning. We further show that step-level scores strongly correlate with final-answer accuracy, highlighting the importance of evaluating reasoning processes in addition to final outcomes.

2 Related Works

2.1 Multimodal Chain-of-Thought Reasoning

Chain-of-Thought (CoT) prompting [44] has achieved strong performance in text-based reasoning, but purely textual rationales provide limited grounding for problems that require spatial perception or diagram understanding. To address this limitation, recent work has explored interleaved multimodal CoT [36, 54], where textual and visual inference are jointly used across intermediate reasoning steps.

Existing explicit multimodal CoT methods mainly follow two directions. The first is tool-augmented reasoning, which invokes external visual modules, drawing tools, or code interpreters to support intermediate reasoning [21, 51, 17, 12, 57, 25]. While effective in some settings, such pipelines are often sensitive to tool interfaces and error propagation, limiting their flexibility and robustness. The second direction studies intrinsic visual CoT within unified model architectures. For example, THINKMORPH [15] investigates emergent reasoning abilities under multimodal interleaved training, while MathCanvas [37] develops an intrinsic VCoT framework for mathematical reasoning. However, most text-to-image models are pre-trained on open-domain visual data for general-purpose generation, limiting their effectiveness for structured visual reasoning. V-CAST [28] improves video inference efficiency through spatio-temporal token pruning. NPS [10], GraftLLM [11], and DiDi-Merging [9] study task-vector compression, modular knowledge fusion, and compact dynamic model merging, respectively.

2.2 Multimodal Reasoning Dataset

High-quality multimodal reasoning data remains another key bottleneck. Early multimodal reasoning benchmarks mainly focus on visual understanding, where models are required to interpret visual inputs and produce textual answers. Representative datasets include MMMU [50], MathVista [32], ScienceQA [31], and MathVerse [53]. These benchmarks have substantially advanced multimodal model evaluation, but their solution processes are often text-centric and provide limited supervision for intermediate cross-modal reasoning. More recent unified benchmarks, such as Uni-MMMU [58], RealUnify [38], and others [42, 18], evaluate both perception and generation abilities. However, they mainly target foundational multimodal capabilities and typically lack the graduate-level depth and systematic step-level protocols needed to audit complex reasoning traces. Other works, like Zebra-CoT [20], further explore reasoning with visual generation. Nevertheless, existing STEM benchmarks often suffer from *modality redundancy*, where models can exploit textual shortcuts without genuinely relying on visual evidence. Many datasets also remain insufficiently challenging for evaluating deep, multi-step reasoning in state-of-the-art MLLMs.

2.3 Fine-Grained Evaluation and Supervision of Reasoning Processes

Most existing evaluation protocols emphasize final-answer accuracy, offering limited insight into the faithfulness, coherence, and grounding of intermediate reasoning steps. To address this issue, recent work has increasingly focused on process-level supervision. Process Reward Models, such as PRM800K [26] and DREAM-PRM [4], demonstrate the value of step-level feedback for improving reasoning reliability and reducing logical hallucinations. Meanwhile, the LLM-as-a-Judge paradigm, including JudgeLM [56] and Prometheus [19], has become a common framework for evaluating open-ended reasoning. Fine-grained evaluation has also begun to emerge in multimodal generation. Benchmarks such as T2I-CoReBench [24] and Science-T2I [22] evaluate structured generated content beyond holistic image quality. VidForensics-M1 [29] uses verifiable temporal

Table 1: Detailed statistics of the StepSTEM. The table presents the distribution of problems across various domains, the availability of multi-path solutions (the 2nd Solution and the 3rd Solution), and the average number of total, textual, visual steps, and bounding boxes per reference solution.

Domain	# Questions	Multi-Solution Samples		Average Steps per Solution			
		Has 2nd Sol.	Has 3rd Sol.	Total Steps	Text Steps	Image Steps	BBoxes
Engineering	111	34	14	7.6	4.4	3.2	5.1
Physics	94	30	0	7.0	3.7	3.3	4.6
Chemistry	25	3	0	6.2	3.2	3.0	6.1
Mathematics	30	1	0	6.9	3.6	3.3	4.9
Other	23	4	0	7.0	3.6	3.4	5.7
Overall	**283**	**72**	**14**	**7.2**	**4.0**	**3.3**	**5.0**

grounding for video-forgery detection, while OmniVerifier-C1 [30] provides decomposed graded verification for visual generation. Our work builds on this trend, but focuses on interleaved multimodal STEM reasoning. In this setting, evaluation must assess not only local step quality, but also whether the predicted reasoning trajectory aligns with a valid cross-modal solution path.

3 Methodology

3.1 Overview

Our StepSTEM consists of two core components: a carefully curated multimodal STEM dataset (§ 3.2) and a fine-grained evaluation protocol (§ 3.3). For the dataset, we collect 283 graduate-level multimodal STEM reasoning problems and annotate each with one or more ground-truth reference solutions. The overall construction workflow is shown in Figure 1, and the domain distribution of StepSTEM is reported in Figure 2. For evaluation, our framework assesses both final-answer correctness and intermediate reasoning quality, including local step quality, global trajectory alignment, and their fusion. When multiple reference solutions are available, we compare the prediction against each reference independently and report the best score.

3.2 Data Curation

To support process-level evaluation, we curate problems with substantial visual dependency. Specifically, we include two types of samples: (1) *strongly dependent* problems, where the solution cannot be reliably derived without the image; and (2) *reasoning-simplifying* problems, where visual information substantially shortens the reasoning chain or reduces semantic ambiguity. The curation pipeline consists of data collection, data filtering, and solution structuring. This process was conducted by eight experts with interdisciplinary backgrounds, who converted raw multimodal problems into structured reasoning trajectories suitable for fine-grained evaluation. Detailed dataset statistics are in Table 1.

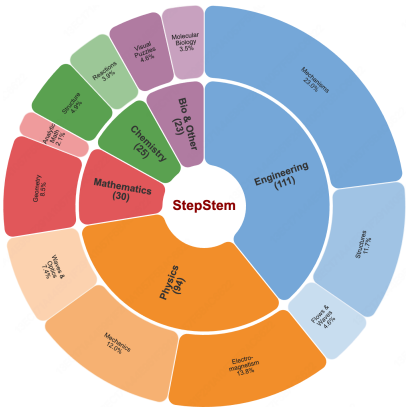

Figure 2: Domain distributions of StepSTEM.

Data Collection Experts collected candidate problems from public competitions, graduate-level textbooks, and other open educational resources. For each candidate, we recorded the problem statement, reference image, ground-truth answer, and a standard solution trajectory containing key intermediate steps. We followed a task-attribute-first principle to assess visual dependency: a problem was retained only if solving it required geometric structures, spatial relations, or local visual details, or if the image substantially simplified the reasoning process. To broaden the evaluation setting, we also include 14 drawing tasks whose final output is an image. These tasks extend the benchmark from conventional "image-text understanding to text generation" toward the more challenging setting of "image-text understanding to image generation."

Data Filtering We applied strict filtering criteria to ensure data quality and evaluation reliability. We removed samples where: (1) the image was purely decorative or redundant with the text; (2) the problem was ambiguous or mainly tested OCR ability; or (3) the image quality was too low to support stable annotation. We further applied a two-sided difficulty filter: problems that experts agreed required no explicit intermediate reasoning were discarded to avoid reducing the benchmark to basic visual recognition. Conversely, problems were also excluded if experts could not reach a stable consensus or formulate a verifiable step-level solution process. Human validation of the difficulty labels is reported in Appendix G. The resulting domain distribution reflects different levels of intrinsic visual dependency across STEM fields. Engineering and Physics naturally contain more graduate-level problems that require spatial or geometric reasoning, whereas **many candidate problems in Mathematics and Biology were removed because they could be solved through purely textual reasoning**. This filtering strategy helps reduce unimodal shortcuts and ensures that retained samples require meaningful use of visual evidence.

Solution Structuring We structure expert-verified solutions into interleaved image-text reasoning trajectories, using GPT-5.4 Codex only for standardized step decomposition and modality-tag assignment under fixed prompts, followed by manual review to ensure each step is independent and verifiable. For visual reasoning steps, experts annotate the relevant local evidence with one or more bounding boxes, enabling fine-grained supervision for step-level image alignment. For problems with multiple valid solution paths, we construct alternative references through expert exploration and LLM-assisted strategy proposal, retaining only those verified by experts as correct, logically consistent, and substantially distinct. As a result, multiple references in STEPSTEM capture genuinely different deductive paths rather than paraphrastic variants, supporting robust *best-over-solutions* evaluation. More details for solution structuring are shown in Appendix A.1.

Through this pipeline, we construct STEPSTEM, a dataset of 283 problems spanning Engineering, Physics, Chemistry, Mathematics, Biology and Other domains.

3.3 Evaluation Protocol

Beyond final-answer accuracy, STEPSTEM evaluates the quality of intermediate multimodal reasoning traces. Each predicted trajectory is compared with one or more annotated reference solutions through two complementary signals. First, a local alignment score measures step-level correspondence: textual steps are matched using attention-rollout-based semantic similarity followed by monotonic dynamic programming, while visual steps are matched by comparing BBox-anchored reference features with sliding-window features from predicted images. A diversity penalty is applied to prevent a single generic generated image from satisfying multiple visual steps. Second, a global LLM judge assesses whether the full predicted trajectory covers the critical textual propositions and visual evidence required by the reference solution. The local and global scores are fused with a weighted RMS, and when multiple valid reference solutions are available, we report the maximum fused score over all references. Final-answer correctness is evaluated separately for textual and visual outputs. More technical details, including textual reasoning step evaluation, visual reasoning step evaluation, global logic coverage evaluation, metric fusion with final answer evaluation, and human evaluation are provided in Appendix A.2. Encoder, judge, alignment, and diversity-penalty sensitivity results are reported in Appendix J.

4 Experiments

4.1 Baselines

To comprehensively evaluate the multimodal reasoning capabilities of current Large Vision-Language Models, we select a diverse set of representative baselines. As outlined in Table 2, these models are categorized into three distinct groups based on their architectural capabilities and accessibility:

Open-Source Unified Models: This group represents the frontier of open-weight models capable of both understanding and generating interleaved text and images. Model details, implementation settings, and prompt templates are provided in Appendices C, B, and N, respectively.

Table 2: Overall results on 269 problems with textual final answer, and 14 problems with image-based final answer. **Txt./Img. Acc.** evaluates the correctness of textual final answers or image-based final answers. **Local (T/I)** evaluates fine-grained alignment of intermediate textual and visual reasoning steps S_{local}. **Global (T/I)** assesses semantic coverage of reference key points and visual regions across the entire trace S_{global}. **Fused** is the integrated process score S_{fused}.

Model	Txt. Acc.	Img. Acc.	Fused	Local-T	Local-I	Local	Global-T	Global-I	Global
Unified Models									
Gemini 2.5 Flash Image	15.61	**50.80**	67.05	58.62	**64.10**	**62.35**	65.57	78.56	71.40
GPT-5.4 + GPT-Image	33.21	39.40	**71.52**	57.72	61.83	60.20	74.46	**88.21**	**80.94**
Interleaved Generation Models									
Bagel	4.83	0.00	47.38	30.64	54.15	44.63	50.34	48.70	49.76
Janus-Pro	0.37	0.00	40.99	24.20	53.20	49.52	38.72	19.16	30.11
OmniGen2	3.35	0.00	45.04	23.58	59.05	42.30	45.97	45.91	47.76
Showo2	1.86	0.00	42.17	35.39	54.13	47.44	41.58	28.33	36.13
Understanding-only Models									
Gemini 3.1 Pro	**38.29**	0.00	38.13	**66.13**	0.00	36.65	71.49	0.00	39.53
Claude Opus 4.6	**38.29**	0.00	37.87	61.38	0.00	34.15	**74.98**	0.00	41.25
Qwen3.5-122B	34.94	0.00	36.55	64.72	0.00	35.45	68.78	0.00	37.43
Qwen3.5-Plus	29.00	0.00	35.01	56.62	0.00	31.05	70.46	0.00	38.46
Qwen3-VL-235B	23.42	0.00	33.18	64.65	0.00	34.26	58.61	0.00	32.05

Table 3: Overall results under the split of correctness. For example, "81.43 / 64.73" indicates the averaged fused score for correct testing examples is 81.43, while the fused score for the incorrect testing examples is 64.73.

Model	Fused	Local-T	Local-I	Local	Global-T	Global-I	Global
Unified Models							
Gemini 2.5 Flash	81.43 / 64.73	60.46 / 58.29	68.74 / 63.79	65.21 / 61.59	88.29 / 62.14	100.00 / 75.70	95.21 / 68.11
GPT-5.4 + Image	78.41 / 68.81	63.34 / 54.79	68.28 / 58.60	62.78 / 58.92	87.21 / 68.84	95.69 / 87.69	91.13 / 77.58
Interleaved Generation Models							
Bagel	65.84 / 47.46	32.35 / 30.98	67.25 / 54.26	51.80 / 45.00	76.92 / 49.79	80.13 / 49.38	77.29 / 49.94
Janus-Pro	78.88 / 40.82	49.58 / 24.69	53.94 / 53.09	50.37 / 49.52	100.00 / 38.18	100.00 / 18.18	100.00 / 29.42
OmniGen2	52.17 / 44.80	26.32 / 22.85	58.77 / 59.34	45.17 / 43.45	63.89 / 43.50	50.00 / 45.54	58.07 / 46.04
Showo2	45.15 / 43.07	49.95 / 35.87	48.32 / 55.17	51.00 / 48.32	46.67 / 40.23	20.00 / 29.40	38.33 / 37.15
Understanding-only Models							
Gemini 3.1 Pro	42.86 / 35.33	71.26 / 63.54	0.00 / 0.00	38.42 / 35.71	87.27 / 61.74	0.00 / 0.00	46.97 / 34.37
Claude Opus 4.6	41.25 / 35.98	67.43 / 59.15	0.00 / 0.00	36.02 / 33.49	86.80 / 68.53	0.00 / 0.00	46.44 / 38.39
Qwen3.5-122B	42.39 / 33.63	70.21 / 62.29	0.00 / 0.00	36.85 / 34.78	89.88 / 57.89	0.00 / 0.00	47.33 / 32.34
Qwen3.5-Plus	41.72 / 31.00	68.74 / 51.91	0.00 / 0.00	36.24 / 28.98	88.27 / 60.94	0.00 / 0.00	46.45 / 34.13
Qwen3-VL-235B	41.43 / 31.02	70.60 / 62.72	0.00 / 0.00	37.53 / 33.26	86.08 / 50.60	0.00 / 0.00	45.07 / 28.17

Closed-source Unified Models: This group represents the current state-of-the-art commercial systems capable of native multimodal generation. We evaluate Gemini 2.5 Flash Image [13] and a pipeline approach combining GPT-5.4 with GPT-image [33, 34].

Understanding-only Models: To demonstrate the necessity of unified generation, we also evaluate powerful models restricted to textual output, including Claude Opus 4.6 [2], Gemini 3.1 Pro [14], and various scales of the Qwen family.

4.2 Main Results

Table 2 reports overall performance on 269 problems with textual final answers and 14 problems with image-based final answers. We group the evaluated MLLMs into three categories according to their reasoning paradigms: unified models, interleaved generation models, and understanding-only models. Domain-specific results appear in Appendix D.

The benchmark remains highly challenging. Overall results show that STEPSTEM is difficult even for strong contemporary MLLMs. On the 269 text-answer problems, the best accuracy is only

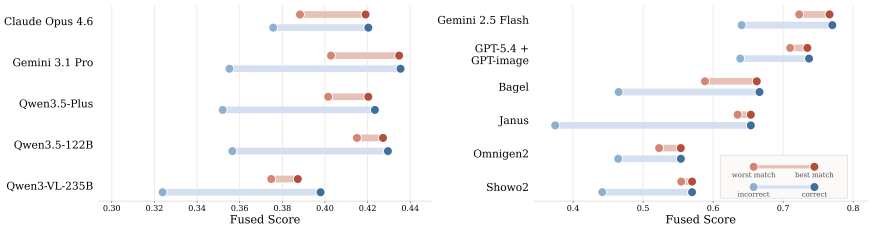

Figure 3: Comparison of the fused process score S_{fused} across different model architectures. The red lines represent the gap between the best-matched and worst-matched ground truth solutions. The blue lines indicate the score difference between trajectories with correct and incorrect final answers.

38.29%, achieved by Gemini 3.1 Pro and Claude Opus 4.6. On the 14 image-answer problems, the highest score is 50.80%, achieved by Gemini 2.5 Flash Image. These results indicate that current MLLMs remain far from reliably solving graduate-level multimodal STEM reasoning tasks.

Vision-text-to-text models exhibit a fundamental limitation. Understanding-only models, such as Claude Opus 4.6 and Gemini 3.1 Pro, achieve relatively strong scores on text-related process metrics, particularly on Local-T and Global-T. However, all such models obtain zero scores on Img. Acc., Local-I, and Global-I. This pattern reveals a fundamental limitation of vision-text-to-text models: while they can process multimodal inputs and produce textual reasoning traces, they cannot handle tasks that require explicit visual generation as part of the reasoning process. Therefore, evaluating only the final answer is insufficient for complex multimodal STEM reasoning, since it fails to capture whether a model can produce and leverage image-based intermediate derivations.

Unified models achieve the strongest multimodal process scores but remain weak at final problem solving. Unified models outperform other categories on fused and global process metrics. GPT-5.4 + GPT-Image achieves the best Fused score of 71.52 and the highest Global score of 80.94, while Gemini 2.5 Flash Image obtains the best Img. Acc. score of 50.80% and a strong Local-I score of 64.10. However, their final-answer accuracy remains limited, particularly on the text-answer subset: GPT-5.4 + GPT-Image reaches 33.21%, while Gemini 2.5 Flash Image reaches only 15.61%. These results suggest that current unified models can generate plausible intermediate visual structures, but still struggle to effectively exploit them for solving challenging STEM reasoning problems.

Interleaved generation alone is insufficient. Interleaved generation models obtain non-trivial image-oriented process scores, indicating some ability to produce intermediate visual content. For example, OmniGen2 achieves 59.05 on Local-I, and Bagel reaches 48.70 on Global-I. However, both models fail to produce correct image-based final answers, yielding 0% accuracy on the image-answer subset. This suggests that simply inserting image generation into the reasoning chain is not enough. Effective multimodal STEM reasoning requires models to generate relevant visual evidence and integrate it coherently into the solution process. Local-I and Global-I measure intermediate visual correspondence and reference coverage, whereas image-answer performance concerns the final submitted image. Relevant intermediate images need not satisfy the geometry, topology, or structural requirements of the final answer, so process scores and final-image results need not improve together. Additional results are provided in Appendix K.

4.3 Analysis and Case Study

4.3.1 Analysis of Step-Level Evaluation Scores

Step-level process scores are strongly associated with final-answer correctness. As shown by the blue intervals in Figure 3, trajectories that lead to correct final answers consistently receive higher fused process scores than those ending with incorrect answers. This pattern is further confirmed by the correctness split in Table 3. For example, the fused score increases from 64.73 to 81.43 for Gemini 2.5 Flash, from 68.81 to 78.41 for GPT-5.4 + GPT-Image, from 47.46 to 65.84 for Bagel, and from 35.33 to 42.86 for Gemini 3.1 Pro when moving from incorrect to correct cases. The consistency of this trend across unified models, interleaved generation models, and understanding-

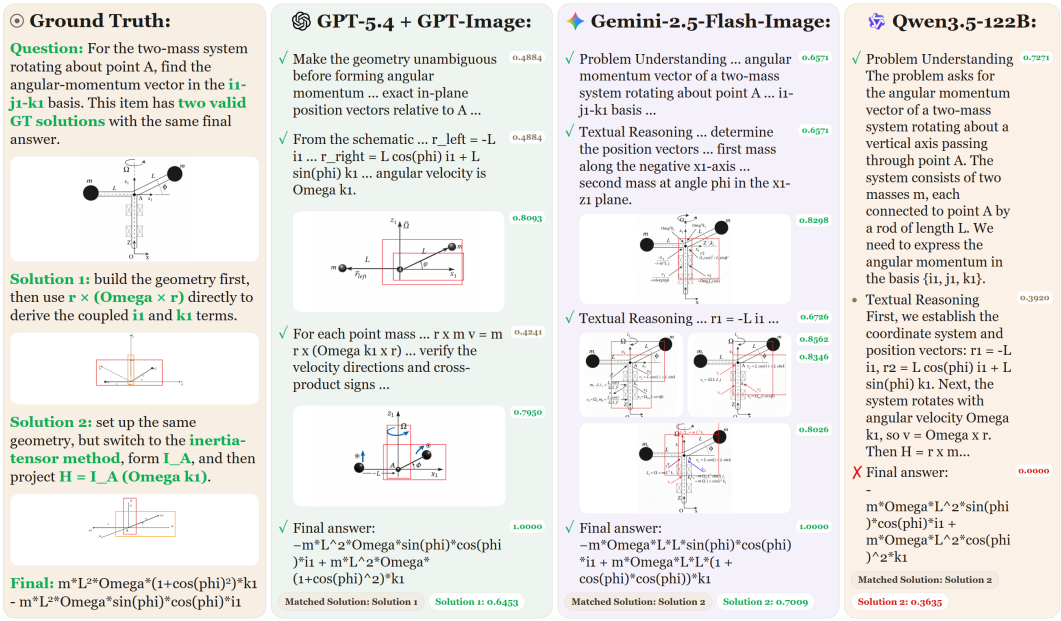

Figure 4: A qualitative case study demonstrating the model's predicted reasoning trajectory versus the ground-truth solutions. Our protocol successfully leverages the multi-solution mechanism and penalizes text-only derivations that lack essential geometric evidence.

only models suggests that the proposed process score captures meaningful variation in reasoning quality rather than incidental overlap with the reference solutions.

Multimodal process coverage complements final-answer accuracy. Beyond the correctness split, Table 2 shows that the two unified models obtain the highest fused scores overall, reflecting their coverage of both textual and visual reference steps. Understanding-only models produce no intermediate images and therefore receive zero Local-I and Global-I under our protocol. Their higher text-answer accuracy need not yield a higher fused score. Figure 3 should thus be interpreted alongside the modality-specific metrics: process coverage and final-answer accuracy provide complementary views of model performance, rather than a single shared model ranking.

The correct–incorrect gap is itself informative. The magnitude of the blue gap in Figure 3 also reveals differences in model behavior. Models such as Bagel and Janus-Pro show relatively large separations between correct and incorrect cases, indicating that successful predictions are supported by substantially better reasoning trajectories. In contrast, weaker models such as Showo2 and OmniGen2 exhibit much smaller gaps, suggesting that their intermediate trajectories remain of limited quality regardless of whether the final answer happens to be correct. This result highlights an important advantage of step-level evaluation: it can distinguish between models that occasionally reach the correct answer and models that more consistently follow a plausible reasoning process.

Overall, these results demonstrate the necessity of evaluating multimodal reasoning at the step level. Final-answer accuracy only indicates whether a model reaches the correct endpoint, but does not reveal whether the model follows a coherent cross-modal solution path. In contrast, our step-level evaluation scores provide a finer-grained diagnostic signal for assessing the quality of intermediate reasoning trajectories. This is particularly important for challenging multimodal STEM problems, where success depends not only on the final prediction, but also on whether the model can sustain a logically consistent reasoning chain across textual and visual steps.

4.3.2 Analysis of Multiple Ground-Truth Solutions

For problems with multiple valid solution strategies, we compute the process score against each ground-truth solution and examine the resulting distribution. As shown by the red intervals in Figure 3, the same model trajectory can receive substantially different scores under different references.

This variation does not necessarily indicate unstable model behavior; rather, it reflects the fact that complex multimodal STEM problems often admit multiple legitimate reasoning paths. A model may therefore follow a correct and coherent solution strategy while still obtaining a low score against one particular annotated reference.

This observation highlights a limitation of single-reference process evaluation. When only one ground-truth solution is available, the metric may unfairly penalize trajectories that are valid but structurally different from the chosen annotation. To mitigate this issue, we provide multiple ground-truth solutions whenever alternative solving strategies exist. With this design, a model can obtain a high process score as long as its reasoning is close to at least one valid reference path. This improves both the stability and the fairness of the evaluation framework by reducing spurious low scores caused by annotation incompleteness.

Moreover, the gap between the best-matched and worst-matched references is itself informative. A larger gap suggests greater diversity in valid solution trajectories and indicates that multi-reference evaluation is particularly important for such problems. Overall, these results show that multiple ground-truth solutions are essential for reliable step-level evaluation of multimodal reasoning. Appendix I gives the single-/multi-reference and best-/worst-match comparisons.

4.3.3 Case Study

Figure 4 shows a representative example involving a two-mass system rotating about point A, where the goal is to derive the angular-momentum vector in the $\mathbf{i}_1$-$\mathbf{j}_1$-$\mathbf{k}_1$ basis. The problem has two valid reference trajectories with the same final answer. Solution 1 directly computes $\mathbf{r} \times (\boldsymbol{\Omega} \times \mathbf{r})$ for each mass, while Solution 2 reformulates the problem using the inertia tensor and derives $\mathbf{H}$ through $\mathbf{I}_A(\Omega\mathbf{k}_1)$. This example illustrates a central challenge in step-level evaluation: correct solutions may follow substantially different intermediate structures. Our framework successfully accommodates such diversity. GPT-5.4 + GPT-Image is aligned with Solution 1, obtaining a matched-solution score of 0.6453 and a perfect final-answer score. In contrast, Gemini 3.1 Pro is aligned with Solution 2, achieving a matched-solution score of 0.7009 while also producing the correct final answer. Thus, models are not forced to mimic the same annotated reasoning chain; instead, they are rewarded when their trajectory is semantically close to any valid reference solution. This demonstrates that our multi-reference evaluation avoids over-penalizing alternative but correct reasoning strategies. The case also highlights the diagnostic value of step-level evaluation for incorrect predictions. Qwen3.5-122B obtains a relatively high problem-understanding score of 0.7271, indicating that it identifies the relevant physical setting and quantities. It also partially aligns with Solution 2 during intermediate reasoning. However, the model fails to recover the complete $\mathbf{k}_1$ component in the final angular-momentum vector, leading to an incorrect final expression and a zero final-answer score. Consequently, its matched-solution score drops to 0.3635. Such a failure pattern would be obscured by final-answer accuracy alone.

Overall, this case study demonstrates two advantages of StepSTEM. First, multi-reference annotation enables fair process evaluation by recognizing distinct but valid reasoning paths. Second, step-level scoring provides a more informative diagnosis of model behavior, revealing whether errors arise from problem misunderstanding, flawed intermediate derivation, or final-step composition. These observations further support the need for fine-grained, multi-reference evaluation in multimodal STEM reasoning.

4.4 Human Evaluation of Reasoning Paths

Figure 5 compares the automatic fused score with the scores from the three human experts. The automatic scores show good qualitative agreement with human annotations. Across all four models, the automatic curve generally follows the same rises and drops as the three human curves, indicating that the proposed protocol captures how far a model can proceed along a valid reasoning path rather than merely measuring superficial overlap with the reference solution. High-scoring trajectories under human evaluation are usually assigned high automatic scores, while trajectories that fail in early reasoning steps receive consistently lower automatic scores. These results indicate that our automatic step-level evaluation is broadly aligned with expert judgement, while also highlighting the importance of multi-solution references and future improvements for handling compressed or semantically equivalent reasoning paths. More results for human evaluations are provided in Appendix E.

Quantitative agreement with expert ratings. Three experts independently evaluated the same 20 problems for four models, yielding 80 responses and 240 individual ratings. Table 4 reports Process Score agreement with mean expert ratings using Pearson, Spearman, Kendall, and mean absolute error (MAE). Process Score has the highest Pearson correlation among the reported metrics for each model. Expert agreement is ICC(2,k)$= 0.8687$ and interval Krippendorff's $\alpha = 0.6866$. Appendix F provides complete metric-by-model results and the remaining inter-annotator statistics; the human scoring protocol is in Appendix A.2.5.

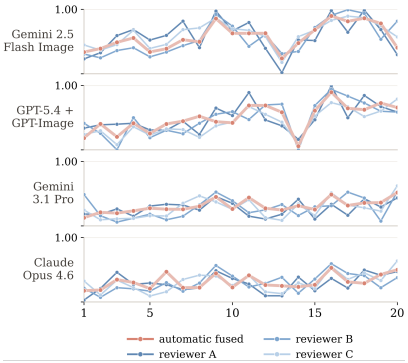

Figure 5: Results for human evaluation.

Table 4: Process Score agreement with mean expert ratings (20 responses per model). MAE uses normalized $[0, 1]$ scores. Full metric-by-model results appear in Appendix F.

Model	Pearson r	Spearman ρ	Kendall τ	MAE
Gemini 2.5 Flash Image	0.7895	0.7614	0.5942	0.1038
GPT-5.4 + GPT-Image	0.8527	0.8249	0.6715	0.0769
Gemini 3.1 Pro	0.8164	0.7927	0.6269	0.0916
Claude Opus 4.6	0.8036	0.7762	0.6118	0.0984

The human target is the mean of three independent ratings of intermediate reasoning progress. Each rating uses the longest correctly covered reference prefix, normalized as m/N, with best-match selection across references. Correlations are computed separately within each model over 20 responses; pooling different model families could obscure differences in output capability. Appendix A.2.5 gives the full scoring procedure.

Table 4 complements the qualitative curves with both association and numerical-error measures. Pearson measures linear association with expert ratings; Spearman and Kendall assess whether response orderings agree, while MAE measures the absolute difference on the shared score scale. Their joint use separates ranking agreement from agreement in score magnitude. Expert agreement concerns the consistency of the human target itself: ICC(2,k) describes the reliability of the averaged ratings, whereas interval Krippendorff's α summarizes agreement among individual ratings. The limited sample and imperfect agreement leave room for evaluation error.

5 Conclusions

We introduced StepSTEM, a benchmark and evaluation framework for interleaved multimodal STEM reasoning. By combining expert-curated problems, step-level textual and visual annotations, and multi-solution trajectory matching, StepSTEM enables evaluation of not only whether a model reaches the correct final answer, but also how it reasons across modalities. Experiments show that current state-of-the-art MLLMs still exhibit a substantial gap between answer correctness and reasoning trajectory quality: understanding-only MLLMs cannot generate visual intermediate steps, whereas unified generative models better expose multimodal reasoning behavior but remain far from reliable. We hope StepSTEM will support future research on multimodal process supervision, reasoning-oriented data construction, and more faithful evaluation of cross-modal reasoning systems.

Coverage and limitations. The 283-problem benchmark is concentrated in Engineering and Physics, with smaller Chemistry and Mathematics subsets and relatively few multi-reference and image-answer problems. Other includes Biology, Materials Science, and Earth Science. Selection for visual dependency further limits the represented task distribution. These coverage constraints should be considered when generalizing model comparisons. Valid unannotated solution paths may still receive low scores. Detailed dataset composition, filtering statistics, and additional limitations appear in Appendices H, L, and M.

References

[1] Samira Abnar and Willem H. Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4190–4197. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.385. URL https://doi.org/10.18653/v1/2020.acl-main.385.

[2] Anthropic. Claude opus 4.6 system card. System card, Anthropic, February 2026. URL https://www-cdn.anthropic.com/14e4fb01875d2a69f646fa5e574dea2b1c0ff7b5.pdf.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[4] Qi Cao, Ruiyi Wang, Ruiyi Zhang, Sai Ashish Somayajula, and Pengtao Xie. Dreamprm: Domain-reweighted process reward model for multimodal reasoning. *arXiv preprint arXiv:2505.20241*, 2025.

[5] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

[6] Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with generated images. *arXiv preprint arXiv:2505.22525*, 2025.

[7] Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

[8] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

[9] Guodong Du and Wanyu Lin. Dynamic model merging made slim. In *Advances in Neural Information Processing Systems*, 2026. URL https://arxiv.org/abs/2605.18904.

[10] Guodong Du, Zitao Fang, Jing Li, Junlin Li, Runhua Jiang, Shuyang Yu, Yifei Guo, Yangneng Chen, Sim Kuan Goh, Ho-Kin Tang, Daojing He, Honghai Liu, and Min Zhang. Neural parameter search for slimmer fine-tuned models and better transfer. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32668–32687. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1570. URL https://aclanthology.org/2025.acl-long.1570/.

[11] Guodong Du, Zhuo Li, Xuanning Zhou, Junlin Li, Zesheng Shi, Wanyu Lin, Ho-Kin Tang, Xiucheng Li, Fangming Liu, Wenya Wang, Min Zhang, and Jing Li. Knowledge fusion of large language models via modular SkillPacks. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10006741.

[12] Chengqi Duan, Kaiyue Sun, Rongyao Fang, Manyuan Zhang, Yan Feng, Ying Luo, Yufang Liu, Ke Wang, Peng Pei, Xunliang Cai, et al. Codeplot-cot: Mathematical visual reasoning by thinking with code-driven images. *arXiv preprint arXiv:2510.11718*, 2025.

[13] Google DeepMind. Gemini 2.5 flash - model card. Model card, Google DeepMind, December 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf. Covers Gemini 2.5 Flash and Gemini 2.5 Flash Image ("Nano Banana").

[14] Google DeepMind. Gemini 3.1 pro. Model card, Google DeepMind, February 2026. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf. Official Gemini 3.1 Pro model card.

[15] Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. *arXiv preprint arXiv:2510.27492*, 2025.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

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

[18] Peizhou Huang, Zixuan Zhong, Zhongwei Wan, Donghao Zhou, Samiul Alam, Xin Wang, Zexin Li, Zhihao Dou, Li Zhu, Jing Xiong, et al. Mmdeepresearch-bench: A benchmark for multimodal deep research agents. *arXiv preprint arXiv:2601.12346*, 2026.

[19] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, 2024.

[20] Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, et al. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*, 2025.

[21] Huayang Li, Siheng Li, Deng Cai, Longyue Wang, Lemao Liu, Taro Watanabe, Yujiu Yang, and Shuming Shi. Textbind: Multi-turn interleaved multimodal instruction-following in the wild. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9053–9076, 2024.

[22] Jialuo Li, Wenhao Chai, Xingyu Fu, Haiyang Xu, and Saining Xie. Science-t2i: Addressing scientific illusions in image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2734–2744, 2025.

[23] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14369–14387. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.775. URL https://doi.org/10.18653/v1/2024.acl-long.775.

[24] Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? *arXiv preprint arXiv:2509.03516*, 2025.

[25] Yuying Li, Siyi Qian, Hao Liang, Leqi Zheng, Ruichuan An, Yongzhen Guo, and Wentao Zhang. Capgeo: A caption-assisted approach to geometric reasoning. *arXiv preprint arXiv:2510.09302*, 2025.

[26] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

[27] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

[28] Xinying Lin, Xuyang Liu, Yiyu Wang, Teng Ma, and Wenqi Ren. V-CAST: Video curvature-aware spatio-temporal pruning for efficient video large language models. *arXiv preprint arXiv:2603.27650*, 2026. URL https://arxiv.org/abs/2603.27650.

[29] Bowei Liu, Zheng Lu, Yuhan Bian, Xinchen Zhang, Xingming Shui, Yuesheng Huang, Xuhuan Li, Zihao Liu, Yifan Yang, Jun Zhou, and Xiu Li. VidForensics-M1: Meta-detection reinforcement learning with verifiable temporal grounding for AI-generated video forensics. *arXiv preprint arXiv:2608.11201*, 2026. URL https://arxiv.org/abs/2608.11201.

[30] Bowei Liu, Xinchen Zhang, Xuhuan Li, Tianlin Zhang, Kaian Jiang, Jingwei Liu, Congyang Zhao, Yifei Qian, Junhong Liu, Zhiheng Li, Yujiu Yang, Yang Cai, and Ling Yang. OmniVerifier-C1: Decomposed graded verifier for rule-governed generative world modeling. *arXiv preprint arXiv:2610.22355*, 2026. URL https://arxiv.org/abs/2610.22355.

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

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

[33] OpenAI. Introducing gpt-5.4, March 2026. URL https://openai.com/index/introducing-gpt-5-4/. Official product release page.

[34] OpenAI. Gpt image 1 model, n.d. URL https://developers.openai.com/api/docs/models/gpt-image-1. OpenAI API model documentation.

[35] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://www.alibabacloud.com/blog/602894.

[36] Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: bridging logical gaps with multimodal infillings. *arXiv preprint arXiv:2305.02317*, 2023.

[37] Weikang Shi, Aldrich Yu, Rongyao Fang, Houxing Ren, Ke Wang, Aojun Zhou, Changyao Tian, Xinyu Fu, Yuxuan Hu, Zimu Lu, et al. Mathcanvas: Intrinsic visual chain-of-thought for multimodal mathematical reasoning. *arXiv preprint arXiv:2510.14958*, 2025.

[38] Yang Shi, Yuhao Dong, Yue Ding, Yuran Wang, Xuanyu Zhu, Sheng Zhou, Wenting Liu, Haochen Tian, Rundong Wang, Huanqian Wang, et al. Realunify: Do unified models truly benefit from unification? a comprehensive benchmark. *arXiv preprint arXiv:2509.24897*, 2025.

[39] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

[40] Meituan LongCat Team, Bin Xiao, Chao Wang, Chengjiang Li, Chi Zhang, Chong Peng, Hang Yu, Hao Yang, Haonan Yan, Haoze Sun, et al. Longcat-next: Lexicalizing modalities as discrete tokens. *arXiv preprint arXiv:2603.27538*, 2026.

[41] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

[42] Chenlong Wang, Yuhang Chen, Zhihan Hu, Dongping Chen, Wenhu Chen, Sarah Wiegreffe, and Tianyi Zhou. Quantifying the gap between understanding and generation within unified multimodal models. *arXiv preprint arXiv:2602.02140*, 2026.

[43] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[44] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[45] Jingxuan Wei, Cheng Tan, Qi Chen, Gaowei Wu, Siyuan Li, Zhangyang Gao, Linzhuang Sun, Bihui Yu, and Ruifeng Guo. From words to structured visuals: A benchmark and framework for text-to-diagram generation and editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 13315–13325. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01243. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wei_From_Words_to_Structured_Visuals_A_Benchmark_and_Framework_for_CVPR_2025_paper.html.

[46] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

[47] Shengguang Wu, Fan-Yun Sun, Kaiyue Wen, and Nick Haber. Symmetrical visual contrastive optimization: Aligning vision-language models with minimal contrastive images. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 30284–30297. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.1462/.

[48] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

[49] Guanyu Yao, Qiucheng Wu, Yang Zhang, Zhaowen Wang, Handong Zhao, and Shiyu Chang. Rethinking the text-vision reasoning imbalance in mllms through the lens of training recipes. *arXiv preprint arXiv:2510.22836*, 2025.

[50] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567, 2024.

[51] Chi Zhang, Haibo Qiu, Qiming Zhang, Zhixiong Zeng, Lin Ma, and Jing Zhang. Deepsketcher: Internalizing visual manipulation for multimodal reasoning. *arXiv preprint arXiv:2509.25866*, 2025.

[52] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.

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

[54] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=y1pPWFVfvR.

[55] Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Baobao Chang, and Minjia Zhang. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China,*

November 4-9, 2025, pages 19666–19690. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.995. URL https://doi.org/10.18653/v1/2025.emnlp-main.995.

[56] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*, 2023.

[57] Xuanyu Zhu, Yuhao Dong, Rundong Wang, Yang Shi, Zhipeng Wu, Yinlun Peng, YiFan Zhang, Yihang Lou, Yuanxing Zhang, Ziwei Liu, et al. Vtc-bench: Evaluating agentic multimodal models via compositional visual tool chaining. *arXiv preprint arXiv:2603.15030*, 2026.

[58] Kai Zou, Ziqi Huang, Yuhao Dong, Shulin Tian, Dian Zheng, Hongbo Liu, Jingwen He, Bin Liu, Yu Qiao, and Ziwei Liu. Uni-mmmu: A massive multi-discipline multimodal unified benchmark. *arXiv preprint arXiv:2510.13759*, 2025.

Appendix

Additional validation and analyses appear in Appendices F–L: quantitative human validation, difficulty-label validation, data curation, reference coverage, evaluator sensitivity, failure modes, and dataset considerations.

A Methodology Details

A.1 Solution Structuring

After filtering (§ 3.2), expert-verified solutions were converted into interleaved image-text reasoning trajectories. To improve consistency and efficiency, we used GPT-5.4 Codex with fixed prompts for initial step decomposition and modality-tag assignment. The model was used only to format expert-provided solutions, not to generate new answers. All structured outputs were then manually reviewed and corrected to ensure that each step forms an independent and verifiable reasoning unit.

For visual reasoning steps, experts annotated the corresponding local visual evidence with bounding boxes. When a step depended on multiple visual regions, multiple bounding boxes were allowed. These annotations identify the specific visual evidence used at each step and provide fine-grained supervision for step-level image alignment evaluation.

For problems admitting multiple valid solution paths, we further constructed alternative references through expert exploration and LLM-assisted proposal. Experts first identified diverse solution strategies based on their domain knowledge. In parallel, we queried advanced multimodal models such as Gemini 3.1 Pro and Claude Opus 4.6 to discover potential strategies that differed from the primary reference solution. All alternative solutions, whether proposed by experts or models, were retained only after expert verification confirmed that they were correct, logically consistent, and substantially distinct from existing references. Thus, multiple references in STEPSTEM represent genuinely different deductive paths rather than surface-level paraphrases, supporting the subsequent *best-over-solutions* evaluation.

A.2 Evaluation Protocol

We evaluate multimodal reasoning traces at two complementary levels: *local step alignment*, which measures whether individual predicted steps correspond to annotated reference steps, and *global logic coverage*, which measures whether the overall trajectory covers the essential propositions and visual evidence required by the reference solution. This design allows the evaluation to reward explicit step-by-step agreement while remaining robust to valid alternative derivations that may not follow the reference trajectory verbatim.

For each problem, let $\mathcal{G} = \{g_1, \ldots, g_m\}$ denote a ground-truth reference trajectory and let $\mathcal{P} = \{p_1, \ldots, p_n\}$ denote the predicted trajectory. Each step is assigned a modality tag. We use $\mathcal{G}^T$ and $\mathcal{G}^I$ to denote the textual and visual reference steps, respectively, and $\mathcal{P}^T$ and $\mathcal{P}^I$ to denote the corresponding predicted steps. Textual reference steps are further annotated with key semantic points, while visual reference steps are annotated with one or more bounding boxes (BBoxes) indicating the local regions that provide necessary visual evidence.

A.2.1 Textual Reasoning Step Evaluation

For textual reasoning, we prioritize semantic process coverage over surface-level string similarity. Given a reference textual step $g_i \in \mathcal{G}^T$ and a predicted textual step $p_j \in \mathcal{P}^T$, we compute a pairwise alignment score using attention rollout [1] from Qwen3.5-9B [35]. Let $W^{(l,h)}$ denote the attention matrix of the h-th head in the l-th transformer layer. We first average attention over heads, add the residual connection, and normalize the resulting matrix:

$$\tilde{A}^{(l)} = \mathrm{RowNorm}\left(\frac{1}{H}\sum_{h=1}^{H} W^{(l,h)} + I\right), \tag{1}$$

where H is the number of attention heads and I is the identity matrix. The global attention rollout matrix is then obtained by multiplying the layer-wise matrices:

$$\bar{A} = \prod_{l=1}^{L} \tilde{A}^{(l)}, \tag{2}$$

where L is the number of transformer layers.

To obtain a fine-grained similarity score between g_i and p_j, we aggregate token-level dependencies from the rollout matrix. For each reference token $u \in g_i$, we identify the predicted token $v \in p_j$ with the maximum rollout weight and average these maxima across all reference tokens:

$$M^T[i,j] = \frac{1}{|g_i|}\sum_{u \in g_i} \max_{v \in p_j} \bar{A}_{u,v}. \tag{3}$$

This yields a textual step-similarity matrix M^T.

We then perform monotonic dynamic programming over M^T to find the optimal alignment path $\pi^{T,*}$. The monotonicity constraint permits skipped steps, allowing models to merge or omit trivial intermediate derivations, while preventing crossed alignments that would violate the temporal order of the reasoning chain. The textual local score is defined as the average similarity over the matched pairs in the optimal path:

$$S^{\text{text}}_{\text{local}} = \frac{1}{|\pi^{T,*}|}\sum_{(i,j)\in\pi^{T,*}} M^T[i,j]. \tag{4}$$

If no valid textual prediction is available, we set $S^{\text{text}}_{\text{local}} = 0$.

A.2.2 Visual Reasoning Step Evaluation

For visual reasoning steps, such as intermediate diagrams, geometric constructions, plots, or state-transition figures, we evaluate whether the predicted images contain the visual evidence specified by the reference solution. Each reference visual step $g_i \in \mathcal{G}^I$ is associated with expert-annotated BBoxes. We encode the annotated region or regions using a visual encoder $\phi(\cdot)$ and use the resulting representation as the reference visual feature. When multiple BBoxes are annotated for the same step, their features are pooled to form the step-level representation.

Predicted visual steps are not assumed to contain annotations. Therefore, for each predicted image step $p_j \in \mathcal{P}^I$, we construct a set of sliding-window patches $W(p_j)$. The pairwise visual similarity between g_i and p_j is defined as the maximum cosine similarity between the reference visual feature and any patch feature extracted from the predicted image:

$$M^I[i,j] = \max_{w \in W(p_j)} \frac{\phi(g_i)\cdot\phi(w)}{\|\phi(g_i)\|\,\|\phi(w)\|}. \tag{5}$$

This produces the visual step-similarity matrix M^I.

A direct maximum-over-windows formulation may overestimate performance when a model repeatedly generates a single generic visual artifact that partially matches many reference steps. To discourage this degenerate behavior, we introduce a diversity-aware penalty. Let

$$C_{\text{img}} = \frac{1}{|\mathcal{G}^I|}\sum_{i=1}^{|\mathcal{G}^I|} \max_j M^I[i,j] \tag{6}$$

denote the maximum visual coverage over all reference visual steps. We further define the image diversity ratio as

$$D_{\text{img}} = \frac{\left|\left\{\arg\max_j M^I[i,j] \;:\; i = 1, \ldots, |\mathcal{G}^I|\right\}\right|}{|\mathcal{G}^I|}, \tag{7}$$

which measures how many distinct predicted visual steps are selected as best matches for the reference visual steps. The final visual local score is computed as

$$S^{\text{img}}_{\text{local}} = C_{\text{img}}\left(1 - \lambda(1 - D_{\text{img}})\right), \tag{8}$$

where λ controls the strength of the diversity penalty. If a problem requires visual reasoning but the model produces no visual step, we set $S^{\text{img}}_{\text{local}} = 0$.

The unified local process score combines textual and visual local scores according to the modality composition of the reference trajectory:

$$S_{\text{local}} = \rho_T S^{\text{text}}_{\text{local}} + \rho_I S^{\text{img}}_{\text{local}}, \tag{9}$$

where

$$\rho_T = \frac{|\mathcal{G}^T|}{|\mathcal{G}^T| + |\mathcal{G}^I|}, \qquad \rho_I = \frac{|\mathcal{G}^I|}{|\mathcal{G}^T| + |\mathcal{G}^I|}. \tag{10}$$

A.2.3 Global Logic Coverage Evaluation

Local alignment evaluates explicit step-level correspondence, but valid reasoning trajectories may differ substantially in structure from the reference solution. To account for such diversity, we additionally introduce an LLM-based global judge that evaluates whole-trajectory coverage.

For each reference solution, we collect a set of critical reference points $\mathcal{R}$, including textual key points and visual evidence specified by BBoxes. The full predicted trajectory is provided to the judge together with these reference points. The judge determines whether each reference point is semantically entailed, mathematically derived, or visually represented in the predicted trajectory. For each $r \in \mathcal{R}$, the judge outputs a binary coverage label:

$$c(r, \mathcal{P}) \in \{0, 1\}, \tag{11}$$

where 1 indicates that the reference point is covered and 0 indicates that it is missing. The global coverage score is then computed as

$$S_{\text{global}} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} c(r, \mathcal{P}). \tag{12}$$

This coverage-oriented judgment rewards trajectories that contain the necessary reasoning content, even when the local step order or granularity differs from the annotated reference.

A.2.4 Final Answer Evaluation and Metric Fusion

We evaluate final-answer correctness independently from process quality. For problems with textual final answers, Qwen3.5-9B is used as an LLM judge to determine whether the predicted answer is equivalent to the ground-truth answer. For drawing tasks or other tasks whose final answer is visual, we evaluate the generated final image using the same visual matching procedure described above.

For process evaluation, we fuse the local alignment score and the global coverage score using a weighted root mean square:

$$S_{\text{fused}} = \sqrt{\frac{w_s S^2_{\text{local}} + w_j S^2_{\text{global}}}{w_s + w_j}}, \tag{13}$$

where w_s and w_j denote the weights assigned to local alignment and global coverage, respectively. By default, we set $w_s = w_j = 0.5$.

Finally, many STEM problems admit multiple valid solution paths. Let $\mathbb{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_K\}$ denote the set of annotated reference trajectories for a problem. We evaluate the predicted trajectory against each reference trajectory independently and take the best score:

$$S_{\text{final}} = \max_{\mathcal{G}_k \in \mathbb{G}} S_{\text{fused}}(\mathcal{G}_k, \mathcal{P}). \tag{14}$$

This best-over-references strategy reduces penalties for following an annotated alternative solution path, while still requiring the predicted trajectory to align with at least one complete and valid reference solution.

A.2.5 Human Evaluation

To further validate whether the proposed automatic step-level evaluation is consistent with expert judgement, we conduct a human evaluation on four representative models: Gemini 2.5 Flash Image, GPT-5.4 + GPT-Image, Gemini 3.1 Pro, and Claude Opus 4.6. These models cover both unified image-text generation systems and understanding-only systems. We select 20 STEPSTEM examples and ask three human experts to independently evaluate the predicted reasoning trajectory of each model.

Given a reference solution containing N ground-truth reasoning steps, each expert identifies the largest prefix length m such that the model prediction correctly covers the first m ground-truth steps in the intended logical order. The normalized human score is then computed as

$$S_{\text{human}} = \frac{m}{N}.$$

For samples with multiple valid reference solutions, we follow the same best-match principle as the automatic protocol and use the highest score over the available references.

B Implementation Details

To ensure complete reproducibility of our evaluation pipeline, we provide the detail of the hyperparameters and configuration settings used throughout the experiments.

B.1 Model Inference Configuration

All baseline models are evaluated under a strict zero-shot setting. The generation temperature is set to 0.2 with a top-p of 0.9 for low-temperature sampling, and the maximum generation limit is capped at 32k tokens. We apply a unified system prompt enforcing a structured, interleaved output format.

B.2 Local Evaluator Configuration

For textual step alignment, we deploy a local Qwen3.5-9B model and aggregate the attention weights. The temperature parameter τ for attention normalization is set to 0.3. For image step alignment, we utilize a pre-trained ResNet-50 [16] encoder. The diversity penalty hyperparameter λ is set to 0.2.

B.3 Global Judge and Metric Fusion

The LLM-based global judge for assessing whole-trace coverage is also powered by Qwen3.5-9B. To reduce sampling variability, we configure it with a temperature of 0.0, a top-p of 1.0, and a maximum token length of 512. In the calculation of the fused process score S_{fused}, the Local and Global scores are assigned equal importance ($w_s = 0.5$ and $w_j = 0.5$).

C Baseline MLLMs

C.1 Closed-source Unified Models

GPT-5.4 + GPT-Image [33, 34] A pipeline approach that explicitly couples OpenAI's advanced large language model (GPT-5.4) with its image generation counterpart (GPT-Image) to handle complex interleaved text-and-image reasoning tasks.

Gemini 2.5 Flash Image [13] A leading commercial multimodal model from Google designed with native capabilities to seamlessly process and generate both textual and visual outputs within a unified inference process.

C.2 Open-Source Unified / Interleaved Generation Models

Janus-Pro [5] An advanced open-source model designed for unified multimodal understanding and generation, operating over both visual and textual modalities in a single framework.

OmniGen2 [46] An open-source multimodal model that integrates a diffusion-based image decoder with LLM hidden states, enabling context-driven image generation and visual reasoning without compromising the model's core textual capabilities.

Showo2 [48] A native unified multimodal model that innovatively fuses autoregressive modeling (for text and understanding tasks) and flow matching/diffusion modeling (for visual generation) within a single Transformer architecture.

Bagel [8] An open-source foundation model featuring a Mixture-of-Transformer-Experts (MoT) architecture. It dedicates separate expert networks to understanding and generation tasks while sharing a unified self-attention mechanism to efficiently handle interleaved text and image data.

C.3 Understanding-only Models

Claude Opus 4.6 [2] Anthropic's flagship model known for exceptional logical reasoning and visual understanding capabilities; however, it operates strictly within a vision-text-to-text paradigm.

Gemini 3.1 Pro [14] Google's advanced multimodal understanding model. It excels at processing complex multimodal inputs to produce detailed textual reasoning traces but cannot explicitly generate visual evidence.

Qwen Family (Qwen3.5-122B, Qwen3.5-Plus, Qwen3-VL-235B) [35, 3] A powerful series of models developed by the Qwen team. While possessing strong foundational language and vision-language understanding, they are fundamentally constrained to text-only generation and rely purely on textual or symbolic reasoning.

D Detailed Experimental Results on Each Specific Domain

Table 5: Model performance on the Engineering subset which includes 105 problems with textual final answer and 6 problems with image-based final answer.

Model	Acc.	Img. Acc.	Fused	Local-T	Local-I	Local	Global-T	Global-I	Global
Unified Models									
Gemini 2.5 Flash Image	11.42	69.55	68.57	58.47	78.70	68.57	62.83	82.65	70.58
GPT-5.4 + GPT-Image	33.20	55.02	69.55	57.44	59.72	57.69	73.04	89.32	79.34
Interleaved Generation Models									
Bagel	1.90	0.00	43.72	28.99	52.57	41.46	47.01	44.07	45.77
Janus-Pro	0.00	0.00	33.37	21.09	47.30	43.70	28.51	11.24	21.32
OmniGen2	0.96	0.00	43.89	26.52	60.44	41.74	45.07	38.90	45.88
Showo2	0.00	0.00	39.69	37.03	56.93	48.56	34.19	24.56	30.85
Understanding-only Models									
Gemini 3.1 Pro	33.33	0.00	40.67	66.22	0.00	38.34	74.86	0.00	42.94
Claude Opus 4.6	31.43	0.00	38.66	61.01	0.00	35.49	73.72	0.00	41.92
Qwen3.5-122B	23.81	0.00	36.06	64.12	0.00	36.81	63.33	0.00	35.22
Qwen3.5-Plus	21.90	0.00	35.35	57.10	0.00	32.91	67.22	0.00	37.68
Qwen3-VL-235B	15.96	0.00	32.84	64.05	0.00	35.57	53.97	0.00	30.16

Tables 5–9 present detailed results on Engineering, Physics, Chemistry, Mathematics, and the combined Other subset. Biology is grouped within Other; a separate Biology table is not reported. Overall, the subset-level analysis is consistent with the main findings in the paper: model performance varies substantially across disciplines, and final-answer accuracy alone does not fully reflect multimodal reasoning ability.

A first notable pattern is the consistent advantage of unified models on image-involved reasoning. On Engineering, Physics, and Chemistry, Gemini 2.5 Flash Image and GPT-5.4 + GPT-Image achieve the strongest image-side process scores. This result indicates that tightly coupling visual understanding and generation is beneficial for problems that require intermediate visual derivations or image-based final outputs. In contrast, understanding-only models remain highly competitive on subsets that favor textual or symbolic reasoning. This trend is particularly clear on Mathematics

Table 6: Model performance on the Physics subset which includes 93 problems with textual final answer and 1 problem with image-based final answer.

Model	Acc.	Img. Acc.	Fused	Local-T	Local-I	Local	Global-T	Global-I	Global
Unified Models									
Gemini 2.5 Flash Image	9.68	0.00	70.05	61.08	66.60	64.66	70.18	80.59	75.16
GPT-5.4 + GPT-Image	24.63	74.18	77.36	61.58	69.62	66.74	77.98	94.96	86.45
Interleaved Generation Models									
Bagel	3.23	0.00	48.67	30.01	52.66	43.89	53.21	52.93	53.16
Janus-Pro	0.00	0.00	42.85	30.41	56.75	53.97	36.01	16.31	27.30
OmniGen2	2.15	0.00	44.66	27.57	58.76	45.90	40.80	44.15	43.51
Showo2	4.30	0.00	41.95	33.67	49.93	44.73	47.48	28.19	38.82
Understanding-only Models									
Gemini 3.1 Pro	33.33	0.00	37.82	68.34	0.00	37.30	70.25	0.00	38.41
Claude Opus 4.6	35.48	0.00	38.44	63.05	0.00	34.57	77.02	0.00	41.81
Qwen3.5-122B	35.48	0.00	38.00	67.39	0.00	36.05	73.76	0.00	39.98
Qwen3.5-Plus	24.73	0.00	35.44	55.82	0.00	29.77	73.81	0.00	40.21
Qwen3-VL-235B	23.78	0.00	34.59	67.32	0.00	34.84	62.85	0.00	34.23

Table 7: Model performance on the Chemistry subset which includes 23 problems with textual final answer and 2 problems with image-based final answer.

Model	Acc.	Img. Acc.	Fused	Local-T	Local-I	Local	Global-T	Global-I	Global
Unified Models									
Gemini 2.5 Flash Image	21.74	59.06	67.02	55.21	66.72	62.72	56.93	82.00	70.98
GPT-5.4 + GPT-Image	30.31	73.68	68.94	53.39	55.35	56.64	65.72	86.98	78.90
Interleaved Generation Models									
Bagel	4.35	0.00	40.61	26.89	43.21	39.00	41.73	42.33	42.02
Janus-Pro	4.35	0.00	48.97	21.06	58.41	52.02	50.00	39.67	45.82
OmniGen2	8.70	0.00	44.61	19.11	58.00	43.04	39.60	53.33	46.13
Showo2	0.00	0.00	43.52	32.69	55.04	46.50	49.60	26.00	40.45
Understanding-only Models									
Gemini 3.1 Pro	43.48	0.00	35.44	63.41	0.00	33.87	68.80	0.00	36.86
Claude Opus 4.6	39.13	0.00	34.80	56.59	0.00	30.29	72.27	0.00	38.85
Qwen3.5-122B	30.43	0.00	34.95	61.70	0.00	32.74	68.73	0.00	36.91
Qwen3.5-Plus	30.43	0.00	33.28	53.48	0.00	28.66	69.80	0.00	37.43
Qwen3-VL-235B	20.40	0.00	31.79	61.63	0.00	31.64	58.57	0.00	31.60

Table 8: Model performance on the Mathematics subset which includes 30 problems with textual final answer.

Model	Acc.	Img. Acc.	Fused	Local-T	Local-I	Local	Global-T	Global-I	Global
Unified Models									
Gemini 2.5 Flash Image	36.67	0.00	59.33	65.61	0.00	34.83	74.83	76.94	76.21
GPT-5.4 + GPT-Image	68.70	0.00	72.82	59.53	62.07	62.06	81.71	83.31	81.90
Interleaved Generation Models									
Bagel	6.67	0.00	58.01	41.54	68.40	58.12	58.28	58.89	57.99
Janus-Pro	0.00	0.00	49.46	25.80	56.31	53.43	64.44	24.72	45.06
OmniGen2	10.00	0.00	51.43	12.03	58.20	35.38	70.78	63.89	67.62
Showo2	3.33	0.00	46.33	42.12	56.92	53.69	41.17	33.33	37.42
Understanding-only Models									
Gemini 3.1 Pro	63.33	0.00	38.74	71.34	0.00	37.55	76.61	0.00	40.02
Claude Opus 4.6	70.00	0.00	41.51	68.51	0.00	36.23	87.44	0.00	46.13
Qwen3.5-122B	76.67	0.00	40.53	70.38	0.00	37.19	83.61	0.00	43.71
Qwen3.5-Plus	70.00	0.00	37.54	64.29	0.00	34.00	77.89	0.00	40.64
Qwen3-VL-235B	51.39	0.00	36.86	70.30	0.00	35.94	71.25	0.00	37.43

subset, where models such as Qwen3.5-122B, Qwen3.5-Plus, and Claude Opus 4.6 achieve the best final-answer accuracy despite scoring zero on image-related metrics. These results suggest that when a task can be largely solved through text-based reasoning after extracting key information,

Table 9: Model performance on the Other subset which includes 18 problems with textual final answer and 5 problems with image-based final answer.

Model	Acc.	Img. Acc.	Fused	Local-T	Local-I	Local	Global-T	Global-I	Global
Unified Models									
Gemini 2.5 Flash Image	27.78	35.16	57.60	43.87	64.18	58.37	57.25	48.91	54.16
GPT-5.4 + GPT-Image	22.13	0.00	58.32	45.65	46.89	47.05	67.01	63.02	67.09
Interleaved Generation Models									
Bagel	27.78	0.00	53.26	31.04	61.14	51.43	53.62	47.47	52.74
Janus-Pro	0.00	0.00	50.42	15.14	57.43	51.59	53.26	39.49	47.46
OmniGen2	5.55	0.00	44.32	13.03	55.80	38.51	46.01	55.44	50.05
Showo2	0.00	0.00	48.17	28.66	53.16	45.96	44.93	43.12	44.25
Understanding-only Models									
Gemini 3.1 Pro	44.45	0.00	29.28	52.87	0.00	27.68	56.52	0.00	29.84
Claude Opus 4.6	38.89	0.00	30.37	52.27	0.00	27.44	59.42	0.00	31.99
Qwen3.5-122B	33.34	0.00	29.59	52.56	0.00	27.17	55.44	0.00	30.10
Qwen3.5-Plus	22.22	0.00	30.18	50.95	0.00	26.08	63.41	0.00	33.36
Qwen3-VL-235B	22.35	0.00	25.82	52.50	0.00	26.26	47.24	0.00	25.77

stronger language reasoning may outweigh the absence of image generation. At the same time, their consistently zero image-answer accuracy highlights a structural limitation of this model family on tasks whose reasoning process or final output must be expressed visually.

E Detailed Experimental Results on Human Evaluation

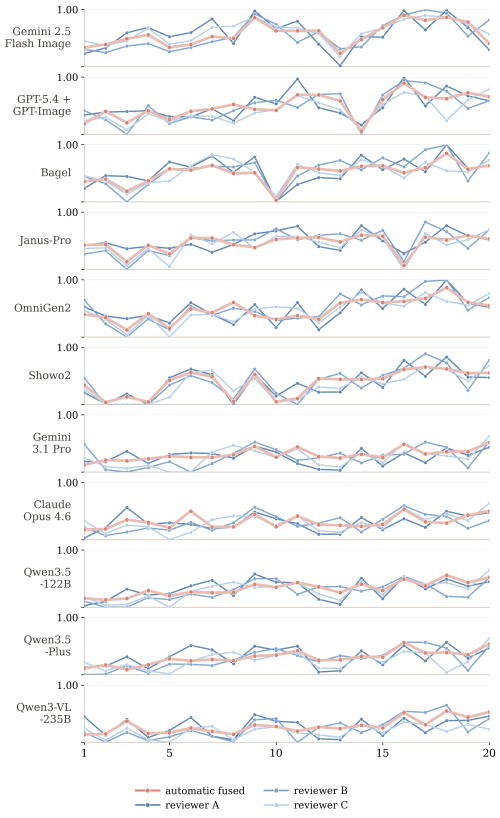

Figure 6: Human evaluation results across all evaluated models. The automatic fused score is compared with the scores assigned by three human reviewers over the same set of examples.

Figure 6 presents qualitative human-evaluation curves for the eleven main-comparison models. The quantitative correlations in Appendix F concern only the four models with supplied summary statistics; they should not be interpreted as validation across all eleven models. Overall, the automatic fused score shows strong qualitative agreement with the scores from the three human reviewers. Across different samples, the automatic curve generally follows the rises and drops of human annotations, indicating that our automatic evaluation is broadly consistent with expert judgement. Meanwhile, the three human reviewers exhibit moderate variation, reflecting the inherent subjectivity of judging complex multimodal STEM reasoning traces. Nevertheless, the automatic scores are generally close to the range of expert scores. These results further validate the proposed step-level automatic evaluation as an effective and scalable proxy for expert assessment of multimodal reasoning processes.

F Quantitative Human Validation of Automatic Metrics

We quantify the agreement between automatic metrics and expert ratings on the same 20 problems and four representative models described in Section A.2.5. For each model, we compare the automatic score for each metric against the mean human rating and report Pearson, Spearman, and Kendall correlations together with mean absolute error (MAE). The human target is the mean of the three independent expert scores, giving 80 model responses and 240 individual ratings. The sample size is 20 for each model–metric pair. Image-metric entries for the two understanding-only models are omitted: these metrics are identically zero under the stated protocol, so their correlations are undefined. Their MAEs require the underlying human ratings and are not reconstructed here. Table 10 reports the complete results.

Table 10: Agreement between automatic metrics and human ratings (20 problems per model).

Model	Metric	Pearson r	Spearman ρ	Kendall τ	MAE
Gemini 2.5 Flash Image	Local-T	0.5624	0.5418	0.3896	0.2147
	Local-I	0.6489	0.6215	0.4673	0.1885
	Global-T	0.5942	0.5761	0.4218	0.2316
	Global-I	0.7168	0.6897	0.5264	0.1572
	Process Score	0.7895	0.7614	0.5942	0.1038
GPT-5.4 + GPT-Image	Local-T	0.7216	0.7043	0.5421	0.1264
	Local-I	0.6847	0.6612	0.5018	0.1496
	Global-T	0.6128	0.5964	0.4487	0.2051
	Global-I	0.7841	0.7556	0.5893	0.1187
	Process Score	0.8527	0.8249	0.6715	0.0769
Gemini 3.1 Pro	Local-T	0.5386	0.5179	0.3742	0.2475
	Local-I	–	–	–	–
	Global-T	0.6672	0.6398	0.4935	0.1867
	Global-I	–	–	–	–
	Process Score	0.8164	0.7927	0.6269	0.0916
Claude Opus 4.6	Local-T	0.5793	0.5534	0.4071	0.2288
	Local-I	–	–	–	–
	Global-T	0.4867	0.4729	0.3415	0.2742
	Global-I	–	–	–	–
	Process Score	0.8036	0.7762	0.6118	0.0984

Process Score has the largest Pearson correlation with human ratings among the reported metrics for each of the four models (0.7895–0.8527). For the two image-generation models, Global-I also correlates more strongly than Local-I, consistent with experts rewarding coverage of visual evidence across the full trajectory. These results support the use of the fused score as a scalable proxy, while the modest sample size and non-perfect correlations leave room for evaluator error.

Inter-annotator agreement is summarized in Table 11. The average pairwise Pearson and Spearman correlations are 0.6886 and 0.6609, respectively. ICC(2,1) is 0.6881 and ICC(2,k) is 0.8687; interval Krippendorff's α is 0.6866.

Table 11: Agreement among the three human annotators.

Statistic	Value
Mean pairwise Pearson / Spearman	0.6886 / 0.6609
ICC(2,1) / ICC(2,k)	0.6881 / 0.8687
Krippendorff's α (interval)	0.6866

G Human Validation of Difficulty Labels

Three domain experts independently assigned difficulty labels to a stratified sample of 50 problems. Majority vote defined the expert-consensus label. The confusion matrix in Table 12 gives exact agreement on 39 of 50 problems (78%). All 50 pairs of consensus and model-derived labels are within one ordinal level (100%). Unweighted Cohen's κ, recomputed from this matrix, is 0.6253. These quantities describe consensus-versus-model agreement, rather than agreement among the three individual experts.

Table 12: Confusion matrix between expert-consensus and model-derived difficulty labels.

Expert consensus	Model Easy	Model Medium	Model Hard	Total
Easy	5	1	0	6
Medium	2	21	3	26
Hard	0	5	13	18
Total	7	27	16	50

The disagreement is concentrated at adjacent difficulty levels: no expert-consensus Easy item is labeled Hard by the model-derived scheme, and no expert-consensus Hard item is labeled Easy. Taking expert consensus as the target and model-derived labels as predictions, precision/recall/F1 recomputed from the matrix are 0.71/0.83/0.77 for Easy, 0.78/0.81/0.79 for Medium, and 0.81/0.72/0.76 for Hard (macro-average 0.77/0.79/0.78). This validation supports the ordinal signal in the difficulty labels, while the sample size does not establish fine-grained calibration for every discipline.

H Dataset Curation Statistics

H.1 Candidate filtering funnel

The original collection began with 4,137 candidate problems. A first filtering stage retained 566 candidates, and expert curation produced the 283-problem benchmark (6.84% of the initial pool; 14.62 candidates per retained problem). Table 13 gives the reported domain counts.

Table 13: Candidate counts and final retained problems in the original 283-problem release.

Domain	Initial candidates	First-filter survivors	Final
Engineering	1,563	219	111
Physics	1,331	193	94
Chemistry	389	52	25
Mathematics	451	58	30
Biology/Other	403	44	23
Total	4,137	566	283

The reported rejection reasons were: insufficient need for interleaved image-text reasoning (1,347; 32.56% of candidates), solvable from text alone (1,173; 28.35%), below graduate-level difficulty (811; 19.60%), decorative image or OCR-only (394; 9.52%), ambiguous or unstable answer (43; 1.04%), and duplicate or highly similar (86; 2.08%). The first two categories account for 60.91%

of the initial candidate pool, reflecting the benchmark's emphasis on text-vision complementarity. Reported domain retention rates are similar: 7.10% for Engineering, 7.06% for Physics, 6.43% for Chemistry, 6.65% for Mathematics, and 5.71% for Biology/Other. Thus, the larger Engineering and Physics subsets are consistent with the composition of the candidate pool and comparable filtering rates, though the benchmark remains weighted toward visually dependent problems in these fields.

The Biology/Other category mainly contains Biology, with some Materials Science and Earth Science problems. Exact counts within this combined category were not supplied in the rebuttal record and are therefore not inferred here.

I Single- and Multiple-Reference Evaluation

Table 14 compares single-reference and multi-reference subsets on the original 283-problem evaluation. The multi-reference subset has 72 problems and the single-reference subset has 211. Multi-reference problems have higher average Process Scores for the three displayed models. Higher scores alone do not establish a difference in intrinsic problem difficulty, because reference coverage also changes the evaluation. The value of multiple references is instead illustrated by evaluating each multi-reference problem against only its worst-matched reference: the aggregate Process Score falls below the single-reference result for all three models. This demonstrates sensitivity to reference choice. Worst-reference selection is an adverse diagnostic and does not establish how a randomly selected single reference would perform; validity of an under-scored trajectory still requires expert verification.

Table 14: Single-reference, multi-reference best-match, and multi-reference worst-match results on the original 283-problem evaluation. Scores are percentages; dashes indicate values not reported.

Model	Subset / matching	N	Text Acc.	Image Acc.	Local-T	Local-I	Global-T	Global-I	Process
GPT-5.4	Single-reference	211	31.96	36.74	56.82	60.27	71.76	86.30	69.54
	Multi-reference, best	72	36.71	47.21	60.36	66.41	82.36	93.82	77.31
	Multi-reference, worst	72	–	–	41.74	64.16	52.27	86.79	62.43
Qwen3.5-Plus	Single-reference	211	27.02	0.00	55.46	0.00	66.72	0.00	33.46
	Multi-reference, best	72	34.51	0.00	60.02	0.00	81.41	0.00	39.57
	Multi-reference, worst	72	–	–	54.31	0.00	47.59	0.00	30.96
BAGEL	Single-reference	211	4.52	0.00	29.69	52.83	48.31	47.31	45.87
	Multi-reference, best	72	5.71	0.00	33.43	58.03	56.28	52.77	51.81
	Multi-reference, worst	72	–	–	30.04	55.37	23.82	41.67	42.73

J Evaluator and Alignment Robustness

J.1 Visual encoder sensitivity

We replace ResNet-50 with CLIP, DINOv2, and SigLIP for the visual local metric. Table 15 reports the resulting Local-I scores for three models. The ranking is unchanged across all four encoders (GPT-5.4, OmniGen2, BAGEL), although absolute scores vary. This supports the comparative ordering for these models, but does not establish that the encoders capture all structured relations in STEM diagrams.

Table 15: Visual encoder sensitivity for Local-I.

Model	ResNet-50	CLIP	DINOv2	SigLIP
GPT-5.4	61.83	73.36	69.43	68.42
OmniGen2	59.05	71.84	67.92	67.11
BAGEL	54.15	66.22	59.72	59.77

J.2 Global judge sensitivity

Table 16 reports Global-T/Global-I scores under four global judges. GPT-5.4 ranks first on Global-T for all four judges. GPT-5.4 also ranks first on Global-I under all four judges; Qwen3.5-Plus scores zero because it produces no images. The Global-T ordering of Qwen3.5-Plus and BAGEL

changes under GPT-4o-mini, so the complete ranking is not invariant. The results therefore support some robustness of text-side ranking, while visual-side comparisons remain dependent on whether a model produces images and on the selected judge.

Table 16: Global judge sensitivity. Each entry gives Global-T / Global-I.

Model	Llama-3.2-11B-Vision	InternVL3.5-8B	Qwen3.5-9B	GPT-4o-mini
GPT-5.4	85.50 / 78.27	71.76 / 75.43	77.17 / 82.13	68.42 / 72.16
Qwen3.5-Plus	83.24 / 0.00	65.42 / 0.00	70.46 / 0.00	62.15 / 0.00
BAGEL	81.38 / 61.87	42.87 / 31.06	50.34 / 48.70	54.38 / 48.72

J.3 Local step alignment sensitivity

We compare monotonic dynamic programming (DP) with greedy and dynamic time warping (DTW) alignment. Table 17 reports the Local-T difference between correct and incorrect text-answer problems, together with the overall Local-T score. Correct and incorrect groups are formed on the 269 text-answer problems; the overall score includes all 283 problems, including the 14 image-answer problems. The overall column is therefore not a weighted average of only the two displayed groups. DP yields positive correct-minus-incorrect gaps for all three models. Greedy and DTW produce negative gaps for BAGEL in this comparison, indicating that the selected alignment procedure affects whether local scores separate correct from incorrect trajectories. For BAGEL with DP, the correct and incorrect text-answer groups contain 13 and 256 problems, respectively; their Local-T means are 32.35 and 30.98. The mean across the 269 text-answer problems is 31.05, whereas the 14 image-answer problems have mean Local-T 22.86. Including all 283 problems gives 30.64, explaining why the overall value is lower than both text-only subgroup means.

Table 17: Alignment method sensitivity. Correct/Incorrect refer to text-answer problems; Overall includes image-answer problems. Gap is Correct Local-T minus Incorrect Local-T.

Model	Method	Correct	Incorrect	Gap / Overall
GPT-5.4	DP	63.34	54.79	+8.55 / 57.72
	Greedy	37.80	29.19	+8.61 / 30.95
	DTW	32.79	29.84	+2.95 / 30.44
Qwen3.5-Plus	DP	68.74	51.91	+16.83 / 56.62
	Greedy	56.31	49.07	+7.24 / 51.09
	DTW	68.11	51.54	+16.57 / 56.17
BAGEL	DP	32.35	30.98	+1.37 / 30.64
	Greedy	25.83	28.77	-2.94 / 28.60
	DTW	28.14	34.33	-6.19 / 33.96

J.4 Diversity penalty sensitivity

We vary the visual diversity penalty λ and compute correlations with human ratings. Table 18 shows the correlations across the tested values. Spearman correlation ranges from 0.5915 to 0.6576, with the largest value at $\lambda = 0.2$, the default setting. This sweep indicates moderate sensitivity rather than invariance; performance degrades at larger penalty values in this sample.

Table 18: Sensitivity to visual diversity penalty λ.

λ	Spearman ρ	Pearson r	Kendall τ
0.0	0.6368	0.5962	0.5031
0.1	0.6489	0.5991	0.5178
0.2	0.6576	0.6005	0.5266
0.3	0.6084	0.5986	0.4793
0.5	0.5915	0.5896	0.4675

K Failure Mode Analysis

The failure analysis covers the eleven main-comparison systems and an additional Qwen3.5-9B baseline (twelve systems in total). We categorize failures on the 269 text-answer problems into seven types: malformed output format; missing visual output; concept or modeling error; inference or logic error; computation, symbol, or unit error; incomplete reasoning; and conclusion-mapping error. Table 19 gives the reported percentages for each model. Percentages use all 269 problems as the denominator and are not normalized over failed responses. Format and concept failures dominate different model families. The supplied aggregate results do not specify a priority rule or whether category assignments overlap; their row sums should therefore not be interpreted as total error rates.

Table 19: Failure mode percentages over 269 text-answer problems.

Model	Format	Visual	Concept	Inference	Comp./Sym./Unit	Incomplete	Conclusion
Gemini 2.5 Flash Image	36.06	5.58	31.23	1.86	1.86	5.95	1.86
GPT-5.4 + GPT-Image	14.55	3.73	32.09	2.24	4.10	7.46	2.61
BAGEL	14.87	7.81	47.21	1.49	4.09	15.99	3.72
Janus-Pro	94.80	0.37	2.97	0.00	0.37	0.74	0.37
OmniGen2	66.91	2.97	17.47	0.37	1.49	5.95	1.49
Show-o2	72.86	2.23	14.87	0.37	1.49	5.20	1.12
Gemini 3.1 Pro	6.32	4.83	29.74	0.37	2.60	15.99	1.86
Claude Opus 4.6	4.83	5.20	37.92	0.74	1.12	9.29	2.60
Qwen3.5-122B	2.97	4.83	33.83	0.74	2.97	16.36	3.35
Qwen3.5-Plus	24.16	5.58	22.68	0.00	3.35	12.27	2.97
Qwen3-VL-235B	1.49	7.06	44.24	1.49	4.09	14.87	3.35
Qwen3.5-9B	19.33	6.32	43.87	0.37	4.83	11.90	2.97

Interleaved-generation systems have high format-error rates in this evaluation, especially Janus-Pro (94.80%), OmniGen2 (66.91%), and Show-o2 (72.86%). Understanding-oriented systems more often fail at concept or modeling, including Qwen3-VL-235B (44.24%) and Qwen3.5-9B (43.87%). Incomplete reasoning remains material across several models, reaching approximately 15–16% for BAGEL, Gemini 3.1 Pro, Qwen3.5-122B, and Qwen3-VL-235B. These patterns help distinguish output-format limitations from errors in the underlying STEM reasoning.

K.1 Intermediate visual quality and final image answers

Six of the eleven systems in the main comparison have image-generation capability: two unified systems and four interleaved-generation systems (BAGEL, Janus-Pro, OmniGen2, and Show-o2). The nonzero Local-I or Global-I scores of some systems do not imply that their final image answers are correct. Local-I measures similarity between intermediate generated visual steps and reference evidence, while image-answer accuracy checks whether the final image satisfies the problem's required answer form. For example, OmniGen2 has Local-I 59.05 but zero image-answer accuracy in the reported evaluation. A model may therefore create a locally relevant intermediate image yet fail to produce a final diagram with the required geometry, topology, or canonical structure. Likewise, higher process scores for correct than incorrect responses are within-model comparisons; they do not imply that every model with a nonzero intermediate visual score can produce a correct final image.

L Additional Dataset and Evaluation Considerations

L.1 Difficulty distribution and model performance

In the original dataset, difficulty counts were reported as 170 Hard, 85 Medium, and 28 Easy problems (60.1%, 30.0%, and 9.9%, respectively). Easy, Medium, and Hard denote the three difficulty strata used in this analysis. Hard is not defined here as requiring every evaluated model to fail. Table 20 reports monotonically decreasing Process Scores from Easy to Hard for GPT-5.4 + GPT-Image, Qwen3.5-Plus, and BAGEL:

This trend is consistent with the intended ordinal interpretation of the labels. The expert validation in Appendix G provides a small-sample check of agreement between the recorded labels and expert consensus. These labels should not be interpreted as a universal intrinsic difficulty scale.

Table 20: Reported Process Scores by difficulty on the original dataset.

Model	Easy	Medium	Hard
GPT-5.4 + GPT-Image	82.14	74.83	68.12
Qwen3.5-Plus	46.28	38.67	31.32
BAGEL	58.29	51.27	43.64

L.2 Contamination and evaluator selection

The problems were sourced from public textbooks, competitions, and instructional material, so overlap with pretraining corpora cannot be ruled out. The benchmark's requirement for interleaved image-text traces and strict final-answer formatting makes memorizing a final answer alone insufficient for a high process score, but this design does not prove that contamination is absent. This remains a limitation when interpreting model scores.

The local text evaluator and global judge use Qwen3.5-9B because it is open-weight and permits reproducible, cost-controlled inference over many trajectories. Using a shared evaluator also keeps local alignment and global coverage judgments consistent. The global-judge ablation above provides evidence on judge sensitivity, but does not eliminate evaluator bias. ResNet-50 was selected as a lightweight visual encoder; its limitations for diagrams with exact topology, symbols, arrows, and spatial relations remain relevant despite the encoder sensitivity results.

L.3 Expert curation and external validation

The curation team comprised eight contributors: four PhD holders, two master's degree holders, and two undergraduate researchers, with expertise in Physics (2), Chemistry (2), Mathematics (2), Engineering (1), and Biology (1). Problems were assigned by domain, and cross-disciplinary problems were reviewed by two experts. Experts annotated independently during filtering and discussed ambiguous cases; items without stable consensus were excluded. Human evaluation was conducted by three PhD students with backgrounds in Physics, Chemistry, and Engineering, with disagreements resolved by consensus after independent ratings.

The rebuttal proposed an additional audit by an unaffiliated PhD expert, but no completed audit results were included in the supplied material. We therefore describe external re-audit as pending rather than as a completed validation experiment.

M Limitations

The benchmark contains 283 problems. Coverage remains limited and weighted toward visually dependent Engineering and Physics; Chemistry and Mathematics are smaller, and image-answer and multi-reference items are a minority. Expert agreement on 50 problems supports ordinal consistency of the difficulty labels, but not a universal difficulty scale. Public educational sources also leave contamination risk unquantified.

Evaluation uses open-weight, cost-controlled Qwen3.5-9B for text alignment and global judging, but it may miss valid derivations. Attention rollout captures token dependencies rather than semantic entailment. ResNet-50 is lightweight, and four-encoder comparisons preserve the ranking of three models, but feature similarity may miss diagram topology, symbols, arrows, and spatial relations.

Monotonic DP may penalize valid reorderings in DAG-like reasoning; global coverage partly mitigates this. Best-over-reference scoring remains reference-bound and may under-score valid unannotated paths, especially for single-reference items. Human validation is limited in scale, and the proposed external expert audit has no completed results in the supplied rebuttal record.

N Actual Prompt Templates

In this section, we provide the prompt templates used in our experiments for reference and reproducibility. These include the prompts for interleaved generation, which instruct models to combine textual and visual reasoning within a unified inference process, as well as the prompts used in the

LLM-as-a-Judge evaluation framework. We release these templates to improve the transparency of our experimental setup and to support future research on multimodal reasoning and fine-grained evaluation.

1. INTERLEAVED_QUERY_PREFIX

```
You will be given a problem. Please answer by strictly following these instructions:

[Task Requirements]
1. Carefully understand the problem first, then provide a complete step-by-step solution.
2. The solution process must use an "interleaved text-and-visual" format:
- For each key step, first explain the reasoning in text.
- Then provide a "visual illustration" block to support that step.
- Each "Visual Illustration" must be an actual image output generated by the model. It must not be
    ↪ replaced by ASCII art, tables, plain text diagrams, flowcharts written in text, or any other
    ↪ text-only substitute.
3. At the end, provide the final answer separately, and it must be strictly wrapped in:
<final_answer>final answer</final_answer>
4. Do not put any extra explanation inside the `<final_answer>` tag other than the final answer
    ↪ itself.
5. Any requirement in the problem statement about the form of the final submitted answer applies only
    ↪  to the final answer itself, not to the reasoning process before it.
6. Any answer-format requirement in the problem statement constrains only the content inside `<
    ↪ final_answer>...</final_answer>'. It does not constrain the length, level of detail, or
    ↪ structure of the reasoning before it.

[Required Output Format]
- Use the following structure:

Problem Understanding
Textual Reasoning
Visual Illustration
Textual Reasoning
Visual Illustration
...
Final Answer
<final_answer>...</final_answer>

[Style Requirements]
- Start directly with "Problem Understanding"; do not describe user intent or restate instructions.
- Do not output meta narration (e.g., "The user wants...", "I need to...", "Let's...").
- Use only the required section headers above. Do not add extra sections, preface, appendix, or
    ↪ commentary.
- Do not output any headings like "Step 1", "Step 2", etc.
- Fully expand the key reasoning needed to justify the answer. Do not over-compress intermediate
    ↪ derivations.
- Prefer preserving explicit intermediate steps over aggressive summarization.
- Each "Textual Reasoning" block may use multiple sentences when needed to make the reasoning
    ↪ complete and self-contained.
- Each "Visual Illustration" block must be an actual image, not a text block.
- Each required section header should appear on its own line.
- Each "Textual Reasoning" block should be followed by a line break before the next "Visual
    ↪ Illustration" block.
- If the solution naturally contains multiple key substeps, keep them as separate interleaved blocks
    ↪ instead of merging them into one paragraph.
- Each visual illustration must closely match the corresponding step.
- Do not omit key reasoning or derivation steps.

[Hard Constraint]
- <final_answer>...</final_answer> must appear exactly once and be the last non-empty line.
- The last non-empty line must be exactly one block: <final_answer>...</final_answer>.
- Do not output any text after the final answer tag.
- Do not use ASCII art or text-only pseudo-visuals in place of images.
- If information is insufficient, still output <final_answer>insufficient_information</final_answer>.
```

2. TOOL_INTERLEAVED_PROMPT

```
You are solving a multimodal problem.

Please solve it through a genuinely interleaved text-and-image reasoning process. Your goal is not to
    ↪  finish as quickly as possible, but to use generated images as part of the reasoning itself.

Follow these principles:
```

```
1. Produce at least one text reasoning segment.
2. Generate multiple genuinely helpful images and distribute them throughout the reasoning process.
3. Text and images must be interleaved:
   - do not write a long stretch of text first and then dump several images at the end;
   - do not generate multiple images in a row without new reasoning progress between them.
4. The process should look like:
   reasoning -> helpful image -> further reasoning from that image -> another helpful image ->
      ↪ further reasoning ...
5. Every image must directly support reasoning rather than decoration. It may help with:
   - understanding the setup
   - marking key objects
   - highlighting relevant regions
   - clarifying structural relations
   - redrawing a schematic more clearly
   - showing an intermediate analytical result
6. Each image must advance the later reasoning, rather than merely repeat earlier content or serve as
      ↪  a presentation illustration.
7. Before generating an image, explain what is still uncertain, what the image is meant to clarify,
      ↪ and why it is useful.
8. After an image is generated, read it carefully and continue reasoning from what it clarifies.
9. The reasoning must be sufficiently detailed and clear. Do not give only a one-line conclusion.
10. Do not generate images that merely paste large amounts of text on the canvas. Images must provide
      ↪  genuinely useful visual information such as schematics, structural relations, arrows, axes,
      ↪ regions, force directions, connections, or key labels.
11. Any answer-format requirement in the problem statement constrains only the final answer itself.
      ↪ It does not constrain the length, level of detail, or structure of the reasoning before it.
12. Requirements such as "one token", "ASCII only", "no extra text", "one final image only", "exact
      ↪ format", or similar restrictions apply only to the final answer, not to the earlier reasoning
      ↪  process.
13. A bare short answer string by itself is not a valid reasoning process. Before the final answer,
      ↪ provide at least one substantive reasoning block even for easy problems.
14. If the problem includes an image, do not end immediately with only a short final answer unless
      ↪ the reasoning has already been explained in a substantive text block.
15. Only give the final answer when you believe the conclusion is stable.
16. For text-answer problems, the final answer must appear in the final text portion and must end
      ↪ with exactly one <final_answer>...</final_answer> block.
17. Do not output any text after the final answer tag.
18. If the required final answer is itself an image, structure, diagram, conformation, or drawing,
      ↪ then the last generated image should be the answer image itself rather than merely an
      ↪ auxiliary illustration.
19. Even when the final answer is an image, provide at least one substantive text block before the
      ↪ answer image explaining what is being drawn, what key structure or relation it must show, and
      ↪  why it is the correct final answer.
20. For such drawing problems, do not replace the answer image with markdown image syntax, SVG code,
      ↪ a data URI, an attachment placeholder, or any other text surrogate. The final answer should
      ↪ be realized as the last generated image itself.
21. Unless the problem is truly trivial, do not finish before using multiple helpful generated images
      ↪  to advance the reasoning.
```

3. Final-answer Judge Prompts

```
You are a strict and careful judge for final-answer equivalence. Your only job is to decide whether
      ↪ the predicted final answer is truly semantically identical to one acceptable ground-truth
      ↪ answer. Be conservative: when a difference may change the meaning, treat it as not equivalent
      ↪ .

----------------------------------------

You are judging final-answer equivalence.

Task: Decide whether the predicted final answer should count as correct against at least one
      ↪ acceptable ground-truth answer.

Decision policy: Be conservative. Mark the answer correct only when the predicted final answer and a
      ↪ ground-truth answer have the same meaning with no material mismatch. If you are unsure,
      ↪ return 0.

Important rules:
1. Focus on final-answer equivalence only.
2. Ignore harmless formatting differences only: punctuation, whitespace, LaTeX wrappers,
      ↪ capitalization, surrounding filler words, and standard algebraic rewrites.
3. Give credit for mathematically equivalent expressions only when they are fully equivalent, with
      ↪ the same variables, same constraints, same set of solutions, same sign, same units, and same
      ↪ boundaries.
4. Do not give credit when any meaningful content changes, including:
   - different numeric value or approximation that is not clearly intended as the same value
```

```
   - different sign, inequality direction, interval boundary, ordering, or logical relation
   - different variable, symbol, label, object identity, or named entity
   - different chemical structure, substituent, isotope placement, compound name, or mapping between
     ↪ labels such as A/B/C/D/E
   - different biological sequence, gene/protein name, stage, category, or ordering
   - one answer being more vague, more generic, or only partially matching the ground truth
5. For structured answers with multiple items, all items must align correctly. A reordering is
     ↪ acceptable only if the answer is explicitly unordered and every item still matches exactly.
     ↪ If labels or item-to-item correspondences change, return 0.
6. Do not infer correctness from topic similarity. Two answers in the same domain are still wrong if
     ↪ any key term or component differs.
7. Use the candidate response tail only as supporting context when the extracted predicted final
     ↪ answer is awkwardly formatted. Do not use it to invent a better final answer than the
     ↪ extracted one.
8. If the predicted answer is empty, image-only, placeholder-like, or lacks enough information to
     ↪ verify exact equivalence, return 0.
9. Output exactly one block between <judge_result> and </judge_result>.
10. Inside the block output a single JSON object with fields:
   "verdict": 0 or 1
   "matched_gt_index": integer index starting from 1, or 0 if none
   "reason": short string

Checklist before returning 1:
- Same objects or entities?
- Same mapping between labels and items?
- Same mathematical meaning, not just similar form?
- No missing or extra components?
If any answer is "no" or "unclear", return 0.

Problem: {question_text}
Predicted extracted final answer: {pred_block}
Candidate response tail: {raw_block}
Acceptable ground-truth final answers: {gt_block}
```

4. Whole-trace Judge Prompts

```
You are a generous, evidence-based judge for reasoning-process coverage.

----------------------------------------

Task: Given the original problem, a candidate model's whole reply, and a list of reference contents,
     ↪ determine for each reference content whether the candidate reply semantically covers it.

Core principle: Your goal is to judge semantic entailment and coverage, not literal wording overlap.
     ↪ Prefer avoiding false negatives on actually correct reasoning. When the candidate reply is
     ↪ mathematically consistent and clearly on the right solution path, be willing to count
     ↪ semantically implied references as covered. Treat a clearly correct final derivation as
     ↪ positive evidence for intermediate coverage whenever the missing intermediate item is a
     ↪ standard consequence of that derivation.

Judging rules:
1. Match against the entire candidate reply, not only the final answer.
2. Mark 1 if the candidate reply explicitly states the reference content, clearly paraphrases it, or
     ↪ provides enough mathematical or semantic evidence that entails it.
3. Mark 1 if the candidate reply gives an algebraically equivalent formula, an equivalent constraint,
     ↪  an equivalent domain statement, or a stronger statement that clearly subsumes the reference.
4. Mark 1 if the evidence is distributed across multiple nearby or logically connected sentences or
     ↪ equations, as long as together they support the reference content.
5. Mark 1 if a later derivation, equation, or final expression clearly implies the reference content,
     ↪  even if the intermediate wording is omitted.
6. Mark 1 if the candidate reply reaches the same local conclusion or a directly equivalent stronger
     ↪ conclusion by a different but valid derivation path.
7. Do not require the same wording, the same notation, the same variable names, the same derivation
     ↪ order, or the same decomposition granularity as the reference.
8. Ignore harmless differences in formatting, equation rearrangement, symbol renaming, simplification,
     ↪  factorization, unit style, and equivalent notation.
9. If the candidate reply states a more concrete result that logically contains the reference idea,
     ↪ count the reference as covered.
10. Use a generous semantic matching standard for correct or near-correct reasoning traces; do not
     ↪ demand verbatim intermediate steps.
11. If the candidate reply provides the correct equation, invariant, constraint, ranking, or final
     ↪ symbolic result from which the reference naturally follows, count it as covered even when the
     ↪  explicit wording differs.
12. Mark 0 only if the reference content is truly missing, too vague to verify, merely topically
     ↪ related, or contradicted by the candidate reply.
13. Do not give 1 for generic topic overlap without concrete supporting evidence.
```

```
14. When uncertain between 0 and 1, prefer 1 if there is plausible mathematical or semantic evidence
      ↪ in the candidate reply.

Output format:
- Output exactly one block between <judge_result> and </judge_result>.
- Inside the block, output one line per reference in the format step_id=0 or step_id=1.
- Do not output explanations.

Problem: {question_text}
Candidate whole reply: {raw_response}
Reference contents: {ref_block}

----------------------------------------

You are a careful multimodal judge for reasoning-process coverage.

Task: Given the original problem, a candidate model's whole reply, and a GT reference image step,
      ↪ determine whether the candidate reply semantically covers the visual reasoning evidence
      ↪ represented by the reference step.

You will receive multiple images as native model inputs in this exact order: {image_inputs_block}

The important regions are specified by normalized bounding boxes over the GT reference image only.
      ↪ You should compare the GT reference image against the candidate generated images and decide
      ↪ whether the candidate generated images support the same visual reasoning evidence or
      ↪ conclusion associated with this GT step.

Judging rules:
1. Consider both the reference text and the important GT image regions.
2. Mark 1 only if at least one candidate generated image semantically matches the GT visual evidence.
3. Pure text alone is not sufficient for image coverage. If the candidate has no relevant generated
      ↪ image evidence, mark 0.
4. Mark 1 if the candidate image evidence plus the candidate reply together capture the same visual
      ↪ idea, relation, geometry, highlight, or conclusion as the GT reference.
5. Mark 0 if the relevant visual evidence is missing from the candidate image panels, contradicted,
      ↪ or too vague to verify.
6. The question image, if provided, is context only. Do not count it as candidate evidence.
7. When uncertain between 0 and 1, prefer 0 unless there is visible evidence in the candidate image
      ↪ panels.

Output format:
- Output exactly one block between <judge_result> and </judge_result>.
- Inside the block, output exactly one line in the format {step_id}=0 or {step_id}=1.
- Do not output explanations.

Problem: {question_text}
Candidate whole reply: {raw_response}
Reference image step id: {step_id}
Reference text: {reference_text}
Important GT image regions: {bbox_block}
```

5. Score Text-Matching Prompt

```
You are a strict grader.

Your goal is to determine whether the Predicted reasoning step between <predicted> and </predicted>
      ↪ semantically contains the Reference key point between <reference> and </reference>.

Judging procedure:
1. Identify the essential propositions in the Reference key point.
2. Check whether each essential proposition is present in the Predicted reasoning step, allowing
      ↪ paraphrase.
3. If all essential propositions are covered and there is no contradiction, output 1.
4. Otherwise output 0.

Rules:
- Paraphrase counts as match.
- Missing any essential proposition => 0.
- Contradiction => 0.
- Mere topic similarity => 0.
- Extra correct information is acceptable.
- Exactly same => 1.
- Same information => 1.

Predicted reasoning step:
<predicted>
```

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the 283-problem benchmark, expert-curated interleaved reference solutions, and local/global process evaluation. These claims are developed in § 3 and evaluated in § 4, with quantitative human validation in Appendix F.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion summarizes coverage and reference-solution limitations. Appendix M discusses dataset composition, contamination, evaluator bias, visual-feature limitations, monotonic alignment, reference coverage, and the scale of human validation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper introduces an empirical benchmark and evaluation protocol. Its equations define scoring procedures; it does not claim new theoretical results requiring proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: § 3.3 and Appendices A.2, B, and N specify the scoring algorithms, model configurations, generation settings, and exact prompts used to reproduce the evaluation protocol. Benchmark data and core scoring scripts are public; the remaining release-packaging limitations are disclosed in the next item.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Benchmark data and core evaluation scripts are publicly available at `https://github.com/lll-hhh/STEPSTEM`. The current release does not include a complete execution environment or the benchmark index referenced by its example commands, so it does not yet provide a complete, executable reproduction package.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: § 4.1 identifies the evaluated models, while § 3.3 and Appendices A.2, B, and N describe zero-shot inference, generation parameters, evaluator configurations, score fusion, and prompts. No model training is performed for the benchmark evaluation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: The paper reports point estimates on the fixed 283-problem benchmark without confidence intervals, error bars, or variability across repeated independent generations. Expert-rating correlations and inter-rater agreement in Appendix F quantify agreement with human assessment rather than repeated-run uncertainty.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: § 4.1 and Appendix B identify the models and inference settings, using APIs for closed-source models and local inference for open-weight models. Complete GPU specifications, memory requirements, execution times, GPU-hours, and API expenditure are not reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work concerns educational STEM benchmark evaluation and expert annotation, with no release of personal participant records or deployment of a decision system. Expert procedures and limitations are described in § 3.2 and Appendices L and M; gaps in licensing and institutional-review documentation are disclosed in the corresponding checklist items.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The contribution is a benchmark for diagnosing multimodal STEM reasoning, with potential benefits for evaluating model reliability. The work does not introduce a deployed application or a new generative model, and no direct high-risk societal application is identified.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released contribution consists of educational STEM benchmark material and evaluation code. The paper does not release a new pretrained generative model or a dataset identified as presenting a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: Baseline models and evaluation encoders are credited in § 4.1 and Appendix C. Complete per-asset licenses and terms of use, including provenance and redistribution terms for the educational source material, are not documented.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: § 3.2, Table 1, and Appendices A.2 and H document benchmark construction, filtering, annotations, and evaluation. The public repository includes benchmark files, core scoring scripts, and a README; release-completeness and licensing gaps are disclosed in the corresponding items.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: § 3.2 and Appendices E, F, and L describe the eight expert curators, three human reviewers, independent annotation, and consensus procedures. Full participant instructions and compensation details are not reported.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: The manuscript describes expert annotation and human ratings but does not report an institutional review or exemption determination, or participant risk and consent procedures. This answer concerns the missing documentation and does not assert that approval was required or that an approval or exemption was obtained.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: § 3.2 describes LLM assistance with formatting and alternative solution construction under expert review. § 3.3 and Appendices A.2, B, and N describe Qwen3.5-9B for local textual alignment, global coverage assessment, and final-answer judging.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.