MathFigAudit: Subfigure-Level Error Diagnosis for Generated K–12 Mathematical Diagrams

Anonymous ACL submission

Abstract

Text-to-image generation is increasingly used to produce educational materials, including mathematical figures from problem statements. Yet a visually plausible math figure can still contain fine-grained errors: it may violate geometric relations, numerical constraints, coordinates, symbolic labels, or text–image correspondence. We introduce MATHFIGAUDIT, a benchmark for subfigure-level error diagnosis in generated K–12 mathematical figures. The benchmark covers 103 drawing-related knowledge points across primary, junior-high, and senior-high mathematics, with 927 2×2 composite figures and 3,708 subfigure-level samples. Given a composite figure and its corresponding Chinese problem bundle, a model must judge each subfigure independently and describe concrete issues when errors are present. We provide human-audited annotations and an evaluation protocol that separates binary error detection from issue-description alignment. Experiments across closed-source and open-source vision-language models (VLMs) show emerging but incomplete auditing ability. Current systems still miss real defects, over-report normal figures, and often produce diagnoses that do not capture the underlying mathematical inconsistency. A supervised fine-tuning study suggests that MATHFIGAUDIT can provide useful supervision for improving error sensitivity, while leaving substantial room for faithful diagnosis. [1]

1 Introduction

Recent image generation systems have substantially advanced prompt following, text rendering, and image editing, as demonstrated by representative models such as Seedream (Gong et al., 2025; Seedream et al., 2025; ByteDance Seed, 2025), Qwen-Image (Wu et al., 2025), Nano Banana Pro (Google, 2025), and GPT Image (OpenAI, 2025, 2026). These advances are beginning to move educational content production from fully manual figure drawing toward human–AI collaborative generation. For mathematics education, this shift is especially promising: figures can be generated directly from problem statements, reducing the cost of creating textbooks, worksheets, online exercises, and personalized learning materials. Large item banks particularly benefit from this capability, because diverse figure variants are often needed for the same knowledge point, question type, and difficulty level.

However, mathematical figures are not ordinary illustrations. Open-domain images are typically judged by visual realism, clarity, or prompt relevance. In contrast, a math problem figure must satisfy a stricter set of constraints: it must be consistent with the problem statement, preserve the intended mathematical structure, use correct symbolic labels, respect numerical relations, and align each visual element with the corresponding text. A generated figure may look complete and visually plausible while still being mathematically invalid due to incorrect angles, violated point-line relations, misaligned coordinates, or misplaced labels. In this setting, visual plausibility does not imply mathematical correctness.

Figure 1 illustrates representative examples from MATHFIGAUDIT. The generated worksheet figures are visually coherent, but human experts identify fine-grained defects such as misaligned clock hands, unrealistic liquid levels, and incorrect label assignments. These examples motivate a benchmark that audits generated figures as educational artifacts, rather than treating them as reliable inputs for downstream problem solving.

Existing multimodal math benchmarks mostly evaluate problem solving from image-text inputs. MathVista (Lu et al., 2024) tests broad visual mathematical reasoning, MathVerse (Zhang et al., 2024) probes whether models truly use diagrams,

[1] Anonymous code repository: https://anonymous.4open.science/status/MathFigAudit-C304

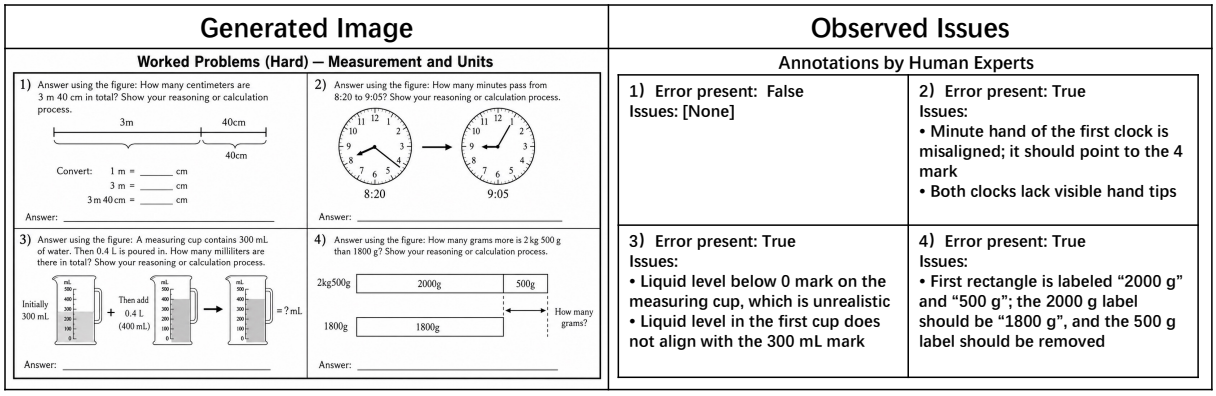


Figure 1: Example of GPT Image 2-generated K–12 mathematical worksheet problems (left) with corresponding observed issues annotated by human experts (right). Each panel shows distinct measurement and unit problems, highlighting fine-grained errors such as misaligned clock hands, liquid level inconsistencies, and incorrect label assignments. This figure illustrates the type of data and evaluation used in the MATHFIGAUDIT benchmark for subfigure-level error diagnosis.

and MATH-Vision (Wang et al., 2024) extends evaluation to competition-level visual math problems. Chart and diagram benchmarks such as FigureQA (Kahou et al., 2017), DVQA (Kafle et al., 2018), and PlotQA (Methani et al., 2020) evaluate reasoning over visualized data, while geometry benchmarks such as GeoQA (Chen et al., 2021), Inter-GPS (Lu et al., 2021), and UniGeo (Chen et al., 2022) combine textual conditions, diagrams, and mathematical reasoning for problem solving. These works generally assume that input figures are reliable, human-designed, and pedagogically valid. In contrast, our setting audits figures automatically generated by text-to-image models, which may look plausible but contain latent defects such as inconsistent geometric relations, incorrect coordinates, distorted proportions, or misplaced labels. It also differs from general text-to-image alignment benchmarks, including TIFA (Hu et al., 2023), GenEval (Ghosh et al., 2023), T2I-CompBench (Huang et al., 2023), and VQAScore (Lin et al., 2024), which mainly assess prompt-level faithfulness, object properties, compositional relations, or image-text alignment. Educational mathematical figures require fine-grained verification of local visual elements against problem constraints. Thus, the model should act as an educational content reviewer: judging figure suitability, localizing defective subfigures, and providing diagnoses for human revision.

We formulate this as a subfigure-level error diagnosis task for generated K–12 math figures. Each input consists of a 2×2 composite figure and a corresponding problem bundle. The model receives the full image once and must output, for each fixed subfigure position, whether the subfigure contains an error and one or more concrete issue descriptions. This composite setting reflects batch-style educational content review, where multiple generated variants under the same knowledge point are inspected side by side. It also introduces a localization and attribution challenge: the model must distinguish similar variants and assign each diagnosis to the correct region.

To support this task, we construct MATHFIGAUDIT, a benchmark covering 103 drawing-related K–12 math knowledge points, 927 composite figures, and 3,708 subfigure-level samples. The dataset is grounded in Chinese K–12 mathematics: the original problem statements and model inputs are in Chinese, while this paper uses English descriptions for exposition. Each subfigure is associated with a binary error label and natural-language issue descriptions derived from human auditing. The errors are not artificially injected template corruptions; they arise naturally from the figure-generation process and include numerical inconsistencies, geometric-structure errors, coordinate or function mismatches, and annotation or notation errors.

Our evaluation separates two abilities that are often conflated. Error detection measures whether a model can identify which subfigures require attention. Issue-description alignment measures whether the model can describe the concrete defect in a way that aligns with human-audited annotations. This distinction is crucial because a model may correctly flag a figure as problematic while failing to identify the cause, or it may produce plau-

sible but non-matching issue descriptions. We evaluate both closed-source and open-source VLMs, analyze composite versus cropped inputs, and include a focused supervised fine-tuning experiment to test whether the dataset can provide useful training signal.

Our contributions are as follows:

- We formulate generated K–12 math-figure auditing as a subfigure-level multimodal diagnosis task, in which a model must decide whether each subfigure is problematic and provide an actionable issue description, going beyond conventional visual math problem solving.
- We construct MATHFIGAUDIT, a benchmark with 927 composite figures and 3,708 subfigure-level samples across 103 drawing-related K–12 math knowledge points, grounded in naturally occurring generation errors rather than synthetic corruption templates.
- We show that current VLMs already exhibit non-trivial auditing ability, but their performance is governed by clear trade-offs among error sensitivity, false-positive control, and issue-description alignment; strong binary detection does not guarantee faithful diagnosis.
- We identify concrete auditing bottlenecks beyond raw visual recognition. Our analyses show that generated math-figure auditing requires both local attribution and mathematical constraint verification, and that broader defect coverage remains easier than producing faithful issue descriptions.
- We provide a focused supervised fine-tuning experiment showing that MATHFIGAUDIT can serve not only as an evaluation benchmark but also as a source of training supervision for improving error sensitivity under a challenging transfer setting.

2 Related Work

Recent multimodal mathematical reasoning benchmarks evaluate whether models can solve problems from image-text inputs. MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2024), MATH-Vision (Wang et al., 2024), WE-MATH (Qiao et al., 2025), MV-MATH (Wang et al., 2025a), and Solid-Geo (Wang et al., 2026) test visual math reasoning at different levels of breadth and difficulty. Chart and geometry benchmarks such as FigureQA (Kahou et al., 2017), DVQA (Kafle et al., 2018), PlotQA (Methani et al., 2020), GeoS (Seo et al., 2015), GeoQA (Chen et al., 2021), Inter-GPS (Lu et al., 2021), and UniGeo (Chen et al., 2022) further require models to interpret structured visual elements. These tasks are primarily solver-centric: the diagram is usually assumed to be correct, and the model is evaluated on the answer, proof, or reasoning trace.

MATHFIGAUDIT is auditor-centric. Instead of asking a model to solve a problem, it asks whether a generated figure is a valid visual representation of the problem and, when it is not, what concrete defect should be fixed. This differs from general image-text faithfulness and text-to-image evaluation, including hallucination and grounding benchmarks (Rohrbach et al., 2018; Thrush et al., 2022; Li et al., 2023; Guan et al., 2024; Wang et al., 2023; Sun et al., 2024; Jing et al., 2024) and generation-evaluation benchmarks or metrics (Heusel et al., 2017; Saharia et al., 2022; Hu et al., 2023; Xu et al., 2023; Ghosh et al., 2023; Huang et al., 2023; Lin et al., 2024). Generated educational math figures require not only prompt faithfulness, but also verification of coordinates, angles, curves, labels, and local mathematical relations.

Educational AI evaluation has studied tutoring, student-response analysis, item generation, feedback, and pedagogical alignment (Graesser et al., 2005; Dzikovska et al., 2013; Gierl and Lai, 2015; Pal Chowdhury et al., 2024; Maurya et al., 2025; Macina et al., 2025; Lelièvre et al., 2025). MATHFIGAUDIT extends this line to generated multimodal teaching materials: it evaluates whether models can support educational content quality control by localizing defective subfigures and producing actionable diagnoses.

3 MATHFIGAUDIT Dataset

3.1 Task and Data Organization

MATHFIGAUDIT encompasses 103 drawing-related K–12 math knowledge points: 13 in elementary school, 20 in junior high, and 70 in senior high. These knowledge points were selected because they inherently require diagrammatic representation, such as triangles, number lines, coordinate systems, functions, analytic geometry, circles, and solid geometry.

For each knowledge point, we generate three

Statistic	Primary	Junior	Senior	Total
Knowledge points	13	20	70	103
Composite figures	117	180	630	927
Subfigures	468	720	2520	3708
Problematic subfigures	53	185	1048	1286
Issue instances	63	239	1452	1754
Avg issues/subfig	1.189	1.292	1.386	1.364

Table 1: Horizontal summary statistics of the MATHFIGAUDIT dataset, including data scale and error counts per educational stage.

question types across three difficulty levels. Each type–difficulty combination contains four generated variants arranged as a 2×2 composite image. Each sample therefore consists of a composite figure I and a problem bundle $Q = \{q_1, q_2, q_3, q_4\}$ with fixed top-left, top-right, bottom-left, and bottom-right subfigure positions. For each subfigure s_i, the model predicts an error label and a list of issue descriptions; the reference annotation follows the same structure. In total, the benchmark includes 927 composite figures and 3,708 subfigure-level samples. Table 1 summarizes the dataset scale and error incidence.

3.2 Generation and Annotation

The construction process follows a human–AI collaborative educational-content pipeline. Human experts select drawing-related knowledge points and revise GPT-4-drafted problem statements to ensure mathematical validity and stage-appropriate difficulty. After the text is confirmed, `gpt-image-2` generates one 16:9 worksheet-style composite image for each type–difficulty group, with four fixed quadrants corresponding to q_1 through q_4. We preserve naturally occurring generation errors instead of injecting template corruptions, so the benchmark reflects realistic generated educational content. Representative dataset examples are shown in Figure 2.

3.3 Annotation Protocol and Quality Control

Each subfigure is annotated with a binary error label and one or more issue descriptions. The binary label indicates whether the subfigure should be considered problematic for the corresponding problem. The issue descriptions record concrete reasons that can guide human revision. Annotators with mathematics-education background inspect each subfigure against the assigned problem, assign a binary error label, and write issue descriptions when an error is present. Disagreements and issue-description conflicts are adjudicated by a senior reviewer, who merges duplicate descriptions, removes unsupported claims, and normalizes descriptions to concrete figure-level defects.

To quantify annotation reliability, we invite two experts with education backgrounds to independently review a sampled subset. Across 848 jointly reviewed subfigures, binary error labels reach 98.5% raw agreement and a Cohen's κ of 0.965. For issue descriptions on jointly positive subfigures, the normalized exact-issue Jaccard overlap is 0.991, indicating that the reviewed labels and concrete issue descriptions are highly stable.

3.4 Analysis of Problem Types in The Generated Dataset

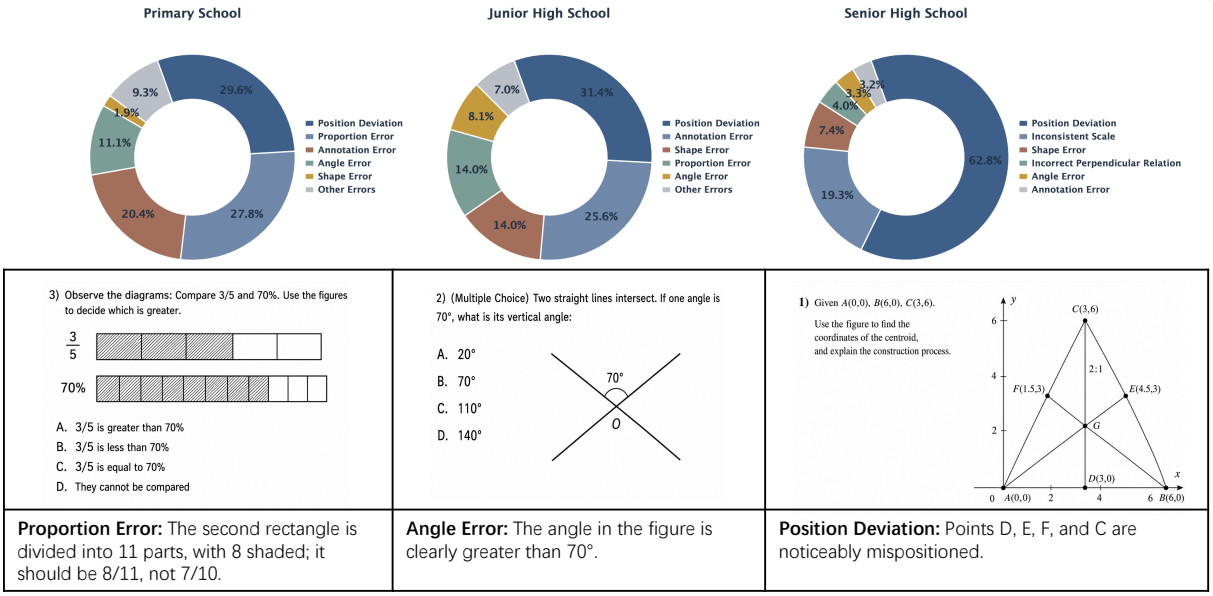
Figure 2 illustrates representative error patterns across educational stages and helps clarify why MATHFIGAUDIT is not a single-skill recognition benchmark. In the primary-school portion, errors often involve visually direct defects such as incorrect proportions, wrong point placement, or mismatched labels. These cases are relatively local, but still require checking whether the depicted structure satisfies the intended quantitative relation in the problem.

The junior-high and senior-high examples are more structurally demanding. Junior-high errors more often involve geometric-shape consistency and auxiliary construction correctness, while senior-high errors frequently depend on coordinate scale, curve shape, or spatial relations that must match stricter mathematical constraints. This progression shows that the generated dataset covers not only superficial drawing flaws but also deeper figure–problem inconsistencies that become increasingly constraint-sensitive at higher educational stages.

4 Experimental Setup

4.1 Input Setting

We consider two input settings. In the composite-image setting, the model receives a full 2×2 image and returns a JSON-style diagnosis for subfigures 1 to 4. In the cropped-subfigure setting, the model receives one cropped subfigure at a time and produces the same JSON-style diagnosis for that subfigure. The output schema is kept consistent across all models and input settings.



Proportion Error: The second rectangle is divided into 11 parts, with 8 shaded; it should be 8/11, not 7/10.	**Angle Error:** The angle in the figure is clearly greater than 70°.	**Position Deviation:** Points D, E, F, and C are noticeably mispositioned.

Figure 2: Examples of common error types in the generated dataset across Primary, Junior High, and Senior High School levels.

4.2 Baselines

We evaluate two categories of VLMs as baselines. The first category consists of closed-source VLMs, including Claude-Sonnet-4.6 (Anthropic, 2026), Gemini-3.1-Flash-Lite-Preview (Google DeepMind, 2026), and GPT-4o (OpenAI, 2024). The second category consists of open-source VLMs. Specifically, we include three Qwen2.5-VL variants, namely Qwen2.5-VL-7B, Qwen2.5-VL-32B, and Qwen2.5-VL-72B (Team, 2025); LLaVA-OneVision-Qwen2-72B (Li et al., 2024); InternVL3-8B and InternVL3-78B (Chen et al., 2024); and InternVL3.5-38B (Wang et al., 2025b). These models cover different model families and parameter scales, allowing us to examine current VLM capabilities for fine-grained educational content auditing.

4.3 Inference and Output Parsing

All models are evaluated with a structured diagnosis prompt that frames the task as figure-quality auditing rather than problem solving; the full prompt template is provided in Appendix Table 7. Decoding uses temperature 0 and top-p 1. Outputs are normalized into a unified JSON-style format by extracting the best-matching structured object and mapping it to the four subfigure entries.

4.4 Metrics

We evaluate each subfigure as a binary auditing decision. Let TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, where positive means that the subfigure contains at least one human-audited issue. The main tables report $\text{Rec.} = \frac{\text{TP}}{\text{TP+FN}}$, $\text{F1} = \frac{\text{2Prec.Rec.}}{\text{Prec.+Rec.}}$, and $\text{Spec.} = \frac{\text{TN}}{\text{TN+FP}}$. Recall measures whether defective subfigures are caught, F1 summarizes the Precision–Recall trade-off, and Specificity measures whether correct subfigures are preserved.

For issue-description alignment, we use a GPT-4 text-only Judge. The Judge is applied on the positive intersection and labels each predicted issue description as `match`, `partial_match`, or `mismatch`; the full Judge prompt and labeling criteria are provided in Appendix Table 8. The main tables report coverage-normalized Judge metrics: $\text{J-Match} = \frac{\#\text{MATCH}}{\#\text{GOLDPOSITIVE}}$ and $\text{J-Partial+} = \frac{\#\text{MATCH}+\#\text{PARTIALMATCH}}{\#\text{GOLDPOSITIVE}}$. Here $\#\text{GOLDPOSITIVE}$ is the number of all human-labeled erroneous subfigures in the evaluated split or stage. Thus, J-Match and J-Partial+ reward a model only when it both detects the erroneous subfigure and provides a semantically aligned diagnosis. We validate the GPT-4 Judge on 200 sampled explanation pairs, obtaining 86.5% accuracy, 0.817 macro-F1, and Cohen's $\kappa = 0.760$ against human labels, with most disagreements between neighboring labels.

Model	Primary					Junior High					Senior High				
	Rec.	F1	Spec.	J-Match	J-Partial+	Rec.	F1	Spec.	J-Match	J-Partial+	Rec.	F1	Spec.	J-Match	J-Partial+
Closed-source VLMs															
Claude-Sonnet-4.6	**0.623**	0.251	0.573	**0.170**	**0.226**	**0.757**	0.491	0.462	0.346	0.438	0.549	0.516	0.539	**0.075**	0.151
Gemini-3.1-Flash-Lite-Preview	0.415	**0.253**	0.761	0.019	0.057	0.697	**0.554**	0.666	**0.416**	**0.524**	**0.594**	**0.532**	0.497	0.056	0.163
GPT-4o	0.132	0.151	0.920	0.038	0.094	0.341	0.403	0.857	0.124	0.205	0.341	0.416	0.766	0.023	0.098
Open-source VLMs															
Qwen2.5-VL-7B	0.057	0.081	0.957	0.000	0.038	0.097	0.147	0.908	0.000	0.027	0.054	0.095	0.929	0.000	0.007
Qwen2.5-VL-32B	0.094	0.114	0.928	0.000	0.019	0.168	0.245	0.919	0.022	0.043	0.161	0.243	0.867	0.017	0.045
Qwen2.5-VL-72B	0.321	0.233	0.817	0.094	0.132	0.276	0.316	0.809	0.108	0.162	0.527	0.515	0.589	0.073	**0.204**
LLaVA-OneVision-Qwen2-72B	0.019	0.034	**0.990**	0.000	0.019	0.005	0.011	**0.996**	0.000	0.000	0.034	0.065	**0.976**	0.001	0.007
InternVL3-8B	0.226	0.203	0.872	0.000	0.057	0.205	0.266	0.862	0.005	0.027	0.309	0.364	0.692	0.012	0.036
InternVL3.5-38B	0.113	0.135	0.928	0.038	0.075	0.189	0.251	0.870	0.011	0.049	0.206	0.283	0.804	0.031	0.063
InternVL3-78B	0.075	0.108	0.959	0.019	0.038	0.243	0.326	0.899	0.027	0.086	0.080	0.139	0.943	0.009	0.017

Table 2: Stage-wise detection and issue-description alignment coverage under the composite-input setting. The best result is shown in bold and the second-best result is underlined for each educational stage and metric.

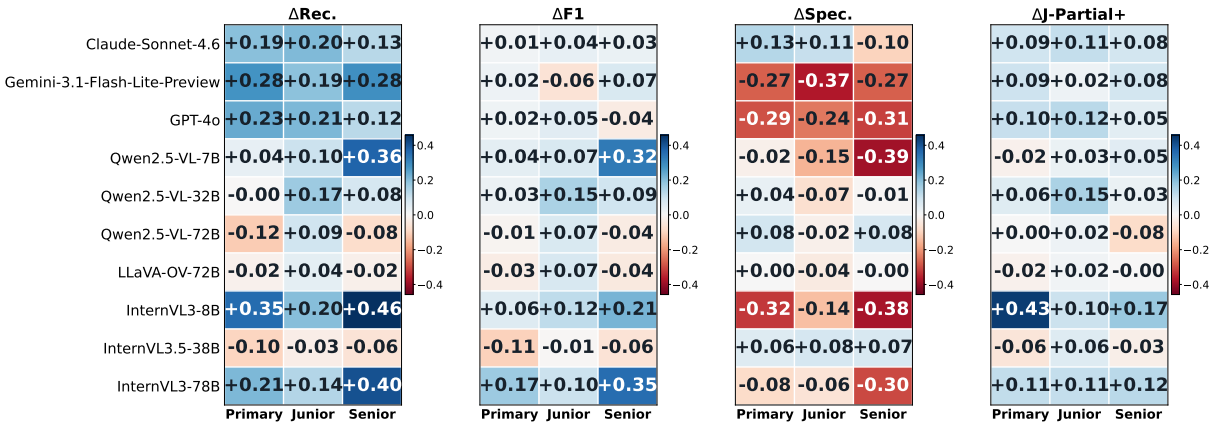


Figure 3: Cropped-minus-composite performance deltas across models and educational stages. Each cell shows the change after replacing the 2×2 composite input with cropped single-subfigure inputs. Blue indicates improvement and red indicates degradation.

5 Results

We organize the results around four guiding questions. The stage-wise benchmark results assess whether current VLMs can act as reliable auditors across primary, junior-high, and senior-high mathematical figures, revealing non-trivial auditing ability alongside clear trade-offs among detection sensitivity, false-positive control, and faithful diagnosis. The subgroup analysis then examines how model behavior varies with question format and nominal difficulty. The cropped-input and error-type analyses further identify the sources of auditing failures: cropping reduces the localization burden for several strong closed-source models, yet the remaining errors indicate that mathematical constraint verification and faithful issue description remain challenging. The supervised fine-tuning experiment finally evaluates whether MATHFIGAUDIT can provide training supervision, with preliminary results suggesting improved error sensitivity on hard transfer splits.

5.1 Main Results: Progress and Remaining Gaps

Current VLMs show emerging ability to audit generated mathematical figures, but their performance reflects clear trade-offs among catching more defective subfigures, avoiding false alarms on correct figures, and providing faithful diagnoses. As shown in Table 2, strong models can detect a substantial fraction of defective subfigures. However, no model simultaneously achieves high Recall, high Specificity, and strong issue-description alignment across all stages. This suggests that MATHFIGAUDIT exposes not a simple visual-recognition problem, but a calibration and diagnosis challenge for educational auditing.

`Claude-Sonnet-4.6` is the most recall-oriented closed-source model on primary and junior-high data, reaching 0.623 and 0.757 Recall, respectively. `Gemini-3.1-Flash-Lite-Preview` provides the strongest overall balance on the denser junior-high and senior-high splits, with F1 scores of 0.554 and 0.532 and a junior-high J-Partial+ of 0.524. Open-source models show larger variance. Conservative

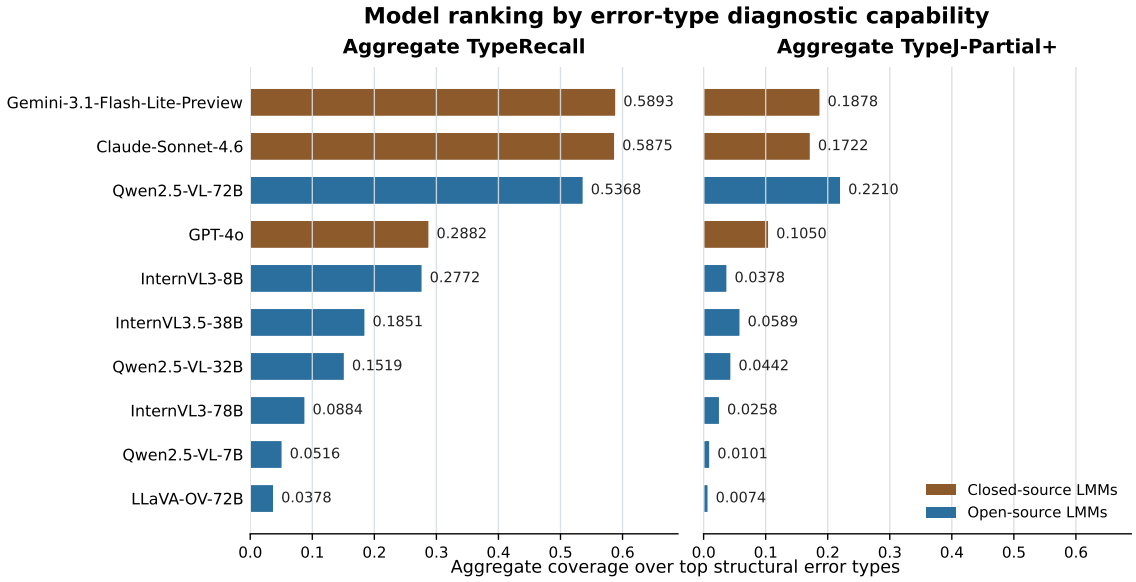


Figure 4: Aggregate diagnostic coverage over common structural error types. The left panel reports aggregate TypeRecall and the right panel reports aggregate TypeJ-Partial+. Models are sorted by TypeRecall. Brown bars indicate closed-source VLMs and blue bars indicate open-source VLMs.

models such as `LLaVA-OneVision-Qwen2-72B` and `Qwen2.5-VL-7B` preserve high Specificity but detect very few actual defects, limiting their use as standalone auditors. In contrast, `Qwen2.5-VL-72B` is the strongest open-source model and remains competitive on senior-high data, where its F1 reaches 0.515, close to `Claude-Sonnet-4.6` (0.516) and `Gemini-3.1-Flash-Lite-Preview` (0.532), although it still lacks stable cross-stage performance.

A central finding is that binary detection and faithful diagnosis are separable abilities. Across stages, J-Match and J-Partial+ remain lower than Recall and F1, indicating that correctly flagging a subfigure does not guarantee that the model has identified the underlying mathematical inconsistency in a way that matches human issue descriptions. For MathFigAudit, this gap is informative rather than merely negative: it shows why generated educational figures require evaluation beyond binary error detection. Position-related errors remain especially prominent in senior-high data, further highlighting that useful auditing requires both sensitivity to defects and actionable diagnosis.

5.2 Diagnostic Analysis of Auditing Bottlenecks

Subgroup analyses show that question format shifts model behavior more strongly than nominal difficulty. Free-response subsets usually induce higher Recall and F1 for strong models, but also lower Specificity, suggesting denser structural defects and more aggressive positive prediction. Difficulty, by contrast, is not monotonic across models: for example, `Claude-Sonnet-4.6` is strongest on medium questions, while `Gemini-3.1-Flash-Lite-Preview` is strongest on easy and hard subsets. Full subgroup results are reported in Appendix B.

Separating Localization from Constraint Verification

Figure 3 summarizes how performance changes when the input is switched from the full 2×2 composite image to cropped single-subfigure images. Cropping removes quadrant attribution and increases Recall for several high-capacity closed models: `Gemini-3.1-Flash-Lite-Preview` gains 0.275, 0.189, and 0.281 on primary, junior-high, and senior-high data, while `Claude-Sonnet-4.6` gains 0.193, 0.201, and 0.126. These gains often come with lower Specificity, and the same trend is not universal for open models: `Qwen2.5-VL-72B` loses Recall on primary and senior-high data (-0.118 and -0.081), while `LLaVA-OneVision-Qwen2-72B` mostly degrades. This indicates that localization is one bottleneck, but mathematical constraint verification remains difficult even after the target subfigure is isolated. Full cropped-input scores and deltas are reported in Appendix Table 9.

Evaluation Split	Setting	Prec.	Rec.	F1	Spec.	J-Partial+
In-topic hard	Base	0.357	0.035	0.065	0.962	0.320
In-topic hard	SFT	0.432	0.113	0.180	0.911	0.381
Held-out-topic hard	Base	0.325	0.031	0.026	0.980	0.213
Held-out-topic hard	SFT	0.444	0.044	0.081	0.951	0.250

Table 3: Supervised fine-tuning results.

Error-Type Diagnosis

The error-type analysis shows that strong models cover a range of recurring structural defect types rather than relying only on dataset-level priors. Figure 4 shows that `Gemini-3.1-Flash-Lite-Preview` and `Claude-Sonnet-4.6` have the highest aggregate TypeRecall, at 0.5893 and 0.5875, followed by `Qwen2.5-VL-72B` at 0.5368. However, TypeJ-Partial+ remains much lower than TypeRecall for all models: the two strongest closed models stay at 0.1878 and 0.1722, while `Qwen2.5-VL-72B` reaches the highest TypeJ-Partial+ at 0.2210. This gap shows that broader defect coverage does not yet translate into faithful issue descriptions. Full type-by-stage heatmaps are provided in Appendix Figures 5 and 6.

5.3 MATHFIGAUDIT as Training Supervision

We conduct a controlled supervised fine-tuning experiment on `Qwen3-VL-32B` using a dedicated difficulty-transfer split derived from MATHFIGAUDIT. We use *in-topic* to denote knowledge points included in the supervised fine-tuning data and *held-out-topic* to denote knowledge points excluded from fine-tuning. The training set contains 374 composite images from 63 in-topic knowledge points and uses only easy and medium samples. Hard samples are excluded from training and split into one validation set and two test sets: the validation set contains 94 hard composite images from in-topic knowledge points, the in-topic hard test split contains another 94 hard composite images from in-topic knowledge points, and the held-out-topic hard test split contains 48 hard composite images from 16 held-out-topic knowledge points. The split unit is the full 2×2 composite image, so all four subfigures from the same image remain in the same split. Additional training details are provided in Appendix C.

The supervised fine-tuning experiment provides evidence that MATHFIGAUDIT can serve not only as an evaluation benchmark, but also as training supervision. The adaptation setting is intentionally challenging: the model is fine-tuned on easy and medium samples, while evaluation is conducted on hard samples from both in-topic and held-out-topic knowledge points. With supervised fine-tuning, F1 rises from 0.065 to 0.180 on the in-topic hard split and from 0.026 to 0.081 on the held-out-topic hard split, with the main change associated with higher Recall. These gains suggest that the benchmark contains learnable signals for improving error sensitivity, including under knowledge-point transfer.

At the same time, the adaptation result remains encouraging but preliminary. Specificity drops on both splits, from 0.962 to 0.911 on the in-topic hard split and from 0.980 to 0.951 on the held-out-topic hard split, suggesting that better detection sensitivity is accompanied by more aggressive positive prediction. The reported J-Partial+ values also remain a supplementary signal here, because they are computed on positive-intersection cases only rather than with the main benchmark denominator over all human-labeled erroneous subfigures. We therefore view this experiment as evidence that MATHFIGAUDIT can help models become more sensitive to figure defects, while calibrated detection and faithful diagnosis remain open challenges.

6 Conclusion

We introduced MATHFIGAUDIT, a benchmark for subfigure-level auditing of generated K–12 mathematical figures, covering 103 drawing-related knowledge points, 927 composite figures, and 3,708 subfigure-level samples. Unlike standard multimodal math benchmarks, MATHFIGAUDIT evaluates whether a model can detect defective generated diagrams and provide a diagnosis that supports human revision. Our experiments show emerging but incomplete auditing ability. Strong VLMs already detect a meaningful portion of defective subfigures, but no model achieves consistently strong Recall, Specificity, and issue-description alignment across settings. The diagnostic analyses further show that generated-figure auditing is limited by multiple bottlenecks, including subfigure localization, mathematical constraint verification, and faithful diagnosis. The focused supervised fine-tuning experiment provides preliminary evidence that MATHFIGAUDIT can also serve as training supervision for improving error sensitivity. We hope MATHFIGAUDIT will support future work on more reliable auditing models, safer educational figure generation, and practical human–AI workflows for mathematical content production.

Limitations

MATHFIGAUDIT is centered on Chinese K–12 mathematics, so its knowledge-point coverage, problem wording, diagram conventions, and naturally occurring error patterns may not transfer directly to other languages, curricula, or figure styles. The benchmark also evaluates issue-description alignment through natural-language descriptions, whose granularity can vary across annotators and predictions; semantic Judge scores help reduce literal mismatch, but issue-description evaluation remains less controlled than binary detection. Finally, the supervised fine-tuning result is intended as a focused proof of training signal under one challenging transfer setting, rather than a comprehensive study of adaptation strategies, model scales, or optimization choices.

Ethical Considerations

This work studies automatic auditing of generated K–12 mathematical figures for educational-content quality control. The dataset is constructed from problem statements and generated diagrams and does not contain student personal data or personally identifiable information. We do not model individual learners, infer sensitive attributes, or make student-level predictions.

The benchmark is intended for research purposes, such as evaluating whether multimodal models can identify diagram defects and provide useful revision-oriented feedback, rather than for high-stakes educational decision making. False positives may increase human review cost, while false negatives may allow flawed figures to enter instructional materials. For this reason, any real-world use should remain under qualified human supervision, and model outputs should be treated as assistive signals rather than final editorial judgments.

As with other data-driven educational benchmarks, the collected error patterns and issue descriptions may reflect biases in the source problems, curricula, annotation practice, or generated-image distribution. These biases may limit transfer to other languages, curricula, diagram conventions, or educational settings. The issue descriptions produced by models are also not guaranteed to be factually correct, pedagogically appropriate, or sufficiently specific for direct classroom use.

Finally, although large language models are used in parts of the pipeline for auxiliary judging and controlled evaluation, they are not treated as authoritative educational experts. Their role is to support benchmark construction and analysis, not to replace teachers, editors, or curriculum specialists. Any downstream deployment should therefore consider the broader educational, social, and ethical implications of automated content auditing.

References

Anthropic. 2026. Introducing Claude Sonnet 4.6. https://www.anthropic.com/claude/sonnet. Released: 2026-02-17; Accessed: 2026-05-19.

ByteDance Seed. 2025. Seedream 5.0 Lite. https://seed.bytedance.com/en/seedream5_0_lite. Accessed: 2026-05-19.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 3313–3323.

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

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Myroslava O Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152.

Mark Gierl and Hollis Lai. 2015. Using automated processes to generate test items and their associated solutions and rationales to support formative feedback. *IxD&A*, 25:9–20.

Lixue Gong, Xiaoxia Hou, Fanshi Li, Liang Li, Xiaochen Lian, Fei Liu, Liyang Liu, Wei Liu, Wei Lu, Yichun Shi, and 1 others. 2025. Seedream 2.0: A

native chinese-english bilingual image generation foundation model. *arXiv preprint arXiv:2503.07703*.

Google. 2025. Nano Banana. https://blog.google/innovation-and-ai/products/nano-banana-pro/. Accessed: 2026-05-19.

Google DeepMind. 2026. Gemini 3.1 Flash-Lite Preview. https://deepmind.google/models/gemini/flash-lite/. Released: 2026-03-03; Accessed: 2026-05-19.

Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48(4):612–618.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Maxime Lelièvre, Amy Waldock, Meng Liu, Natalia Valdés Aspillaga, Alasdair Mackintosh, María José Ogando Portela, Jared Lee, Paul Atherton, Robin AA Ince, and Oliver GB Garrod. 2025. Benchmarking the pedagogical knowledge of large language models. *arXiv preprint arXiv:2506.18710*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pages 23439–23554.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 204–221.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536.

OpenAI. 2024. GPT-4o. https://developers.openai.com/api/docs/models/gpt-4o. Released: 2024-05-14; Accessed: 2026-05-19.

OpenAI. 2025. GPT Image 1.5. https://developers.openai.com/api/docs/models/gpt-image-1.5. Accessed: 2026-05-19.

OpenAI. 2026. GPT Image 2.0. https://developers.openai.com/api/docs/models/gpt-image-2. Accessed: 2026-05-19.

Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 5–15.

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

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, and 1 others. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, and 1 others. 2025. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1466–1476.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.

Qwen Team. 2025. Qwen2.5-vl.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and 1 others. 2023. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

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

Peijie Wang, Zhong-Zhi Li, Fei Yin, Dekang Ran, and Cheng-Lin Liu. 2025a. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19541–19551.

Peijie Wang, Chao Yang, Zhong-Zhi Li, Dekang Ran, Mi Tian, Zhilong Ji, Jinfeng Bai, Cheng-lin Liu, and 1 others. 2026. Solidgeo: Measuring multimodal spatial math reasoning in solid geometry. *Advances in Neural Information Processing Systems*, 38.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, and 20 others. 2025. Qwen-image technical report. *Preprint*, arXiv:2508.02324.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935.

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

A Knowledge Point Coverage

Table 4 lists all 103 knowledge points covered by MATHFIGAUDIT. The Chinese names correspond to the dataset folders and source problem organization, while the English names are provided as readable renderings for this paper.

Table 4: Knowledge points covered by MATHFIGAUDIT, grouped by educational stage.

Stage	Chinese knowledge point	English rendering
Primary	三角形	Triangles
Primary	分数、百分数可视化	Fraction and percentage visualization
Primary	周长与面积	Perimeter and area
Primary	四边形	Quadrilaterals
Primary	圆与组合圆形	Circles and composite circular figures
Primary	展开图与表面积	Nets and surface area
Primary	应用题线段图与关系图	Segment diagrams and relation graphs for word problems
Primary	正比例与反比例图像	Direct and inverse proportion graphs
Primary	比例尺与地图_平面图	Scales, maps, and plan views
Primary	测量与单位图示	Measurement and unit diagrams
Primary	立体图形	Solid figures
Primary	线、角与位置关系	Lines, angles, and positional relations
Primary	轴对称与图形变换	Axial symmetry and geometric transformations
Junior high	几何图形初步(平面-立体-三视图-展开图)	Basic geometric figures: plane figures, solids, three views, and nets
Junior high	直线、射线、线段	Lines, rays, and line segments
Junior high	角	Angles
Junior high	相交线与平行线	Intersecting and parallel lines
Junior high	平面直角坐标系	Cartesian coordinate system
Junior high	三角形	Triangles
Junior high	轴对称	Axial symmetry
Junior high	勾股定理	Pythagorean theorem
Junior high	平行四边形	Parallelograms
Junior high	特殊平行四边形	Special parallelograms
Junior high	一次函数	Linear functions
Junior high	二次函数	Quadratic functions
Junior high	旋转	Rotation
Junior high	圆	Circles
Junior high	概率初步	Introductory probability
Junior high	反比例函数	Inverse proportional functions
Junior high	相似	Similarity
Junior high	锐角三角函数	Trigonometric functions of acute angles
Junior high	投影与视图	Projection and views
Junior high	数据统计图	Statistical charts
Senior high	数轴表示集合	Sets on the number line
Senior high	文氏图表示集合关系	Venn diagrams for set relations
Senior high	区间与不等式解集	Intervals and inequality solution sets
Senior high	一次函数	Linear functions
Senior high	二次函数	Quadratic functions
Senior high	反比例函数	Inverse proportional functions
Senior high	幂函数	Power functions
Senior high	指数函数	Exponential functions
Senior high	对数函数	Logarithmic functions
Senior high	分段函数	Piecewise functions
Senior high	绝对值函数	Absolute value functions
Senior high	函数单调性	Function monotonicity
Senior high	函数奇偶性	Function parity
Senior high	函数周期性	Function periodicity
Senior high	反函数	Inverse functions
Senior high	函数图象变换	Transformations of function graphs
Senior high	弧度制与扇形	Radian measure and sectors
Senior high	单位圆	Unit circle
Senior high	正弦定理	Sine rule
Senior high	余弦定理	Cosine rule

Continued on next page

Stage	Chinese knowledge point	English rendering
Senior high	一元二次不等式	Quadratic inequalities in one variable
Senior high	分式不等式	Rational inequalities
Senior high	高次不等式穿轴法	Polynomial inequalities and sign charts
Senior high	二元一次不等式组	Systems of linear inequalities in two variables
Senior high	线性规划可行域	Feasible regions in linear programming
Senior high	频率分布直方图	Frequency histograms
Senior high	频率分布表	Frequency distribution tables
Senior high	事件关系文氏图	Venn diagrams for event relations
Senior high	古典概型树状图	Tree diagrams for classical probability
Senior high	排列组合分类图	Classification diagrams for permutations and combinations
Senior high	向量概念	Vector concepts
Senior high	共线向量	Collinear vectors
Senior high	向量加法	Vector addition
Senior high	向量减法	Vector subtraction
Senior high	向量坐标表示	Coordinate representation of vectors
Senior high	投影	Projections
Senior high	直线方程	Equations of lines
Senior high	两直线平行	Parallel lines
Senior high	两直线垂直	Perpendicular lines
Senior high	点到直线的距离	Distance from a point to a line
Senior high	线段定比分	Section formula for a line segment
Senior high	三角形重心	Triangle centroid
Senior high	三角形垂心	Triangle orthocenter
Senior high	三角形外心	Triangle circumcenter
Senior high	三角形内心	Triangle incenter
Senior high	对称问题	Symmetry problems
Senior high	轨迹方程	Locus equations
Senior high	圆的方程	Equations of circles
Senior high	直线与圆的位置关系	Positional relations between lines and circles
Senior high	圆的切线	Tangents to circles
Senior high	圆的弦与垂径定理	Chords and the perpendicular diameter theorem
Senior high	椭圆	Ellipses
Senior high	双曲线	Hyperbolas
Senior high	抛物线	Parabolas
Senior high	焦点与准线	Focus and directrix
Senior high	焦点弦	Focal chords
Senior high	通径	Latus rectum
Senior high	中点弦	Midpoint chords
Senior high	直线与圆锥曲线综合	Lines and conic sections
Senior high	空间点线面关系	Spatial relations among points, lines, and planes
Senior high	线线平行与垂直	Parallelism and perpendicularity between lines
Senior high	线面平行与垂直	Parallelism and perpendicularity between lines and planes
Senior high	面面平行与垂直	Parallelism and perpendicularity between planes
Senior high	异面直线所成角	Angles between skew lines
Senior high	直线与平面所成角	Angles between lines and planes
Senior high	二面角	Dihedral angles
Senior high	棱柱	Prisms
Senior high	棱锥	Pyramids
Senior high	球	Spheres
Senior high	球与内接几何体	Spheres and inscribed solids

B Question-Type and Difficulty Breakdowns

Tables 5 and 6 provide the full subgroup results used in the diagnostic discussion. They report the same detection and issue-description alignment metrics as the main table, but aggregate examples by question format and difficulty level instead of by educational stage.

Table 6 further separates the composite-input results by nominal difficulty. This view shows that harder questions do not uniformly reduce model performance; instead, different models shift their Recall, Specificity, and Judge alignment in different ways across easy, medium, and hard subsets.

C Supervised Fine-Tuning Details

The supervised fine-tuning experiment uses the composite-image input setting and the same structured diagnosis target as the benchmark evaluation. Each training instance consists of one 2×2 composite image, four aligned problem texts, and an assistant target in the unified JSON-style schema containing the binary error judgment and issue descriptions for all four subfigures. The SFT data are built from human-audited labels in MATHFIGAUDIT.

We fine-tune `Qwen3-VL-32B` with LoRA rather than full-parameter updating. The LoRA configuration uses rank 64, $\alpha = 128$, and dropout 0.05, with automatically selected common linear projection modules. Optimization uses a learning rate of 1×10^{-5}, 3 epochs, per-device batch size 1, gradient accumulation 16, maximum sequence length 4096, bf16 precision, gradient checkpointing, warmup ratio 0.03, weight decay 0, and random seed 42.

D Prompt and Output Schema

This section provides the output contract used by the input settings and output-parsing procedure described in the main text. All models are prompted to return a JSON-style object with four entries, `subfigure 1` to `subfigure 4`. Each entry contains a Boolean error judgment and a list of concrete issue descriptions. The Boolean field supports subfigure-level detection metrics, while the issue-description field supports the Judge-based semantic-alignment metrics.

E Image-Auditing Prompt for Dataset Construction

This prompt corresponds to the auxiliary auditing step in the dataset-construction pipeline. During dataset construction, closed-source and open-source vision-language models are used as auxiliary auditors to analyze generated figures before human verification. The original prompt is written in Chinese; for readability, we provide an English rendering below while preserving the task definition, constraints, and output schema.

F GPT-4 Judge Prompt and Labeling Criteria

For issue-description alignment evaluation, we use GPT-4 as a text-only Judge. The Judge does not inspect the image or solve the problem; it only compares human-audited issue descriptions with model-predicted issue descriptions.

G Figure Generation Prompt

This section expands the figure-generation procedure summarized in the dataset-construction section. Each knowledge point is expanded into nine type–difficulty groups: easy/medium/hard multiple-choice, fill-in-the-blank, and free-response questions. Each group contains four questions. The image-generation prompt instructs `gpt-image-2` to create one separate 16:9 landscape worksheet-style image for each group. The canvas is divided into four equal quadrants with a fixed order: top-left q_1, top-right q_2, bottom-left q_3, and bottom-right q_4. Each quadrant includes the problem number, a concise problem statement when needed, and the requested mathematical diagram. The style is black-and-white teaching-material style: white background, black lines, clear Chinese typography, no poster-like decoration, and mathematical labels such as point names, angle marks, parallel/perpendicular symbols, coordinate axes, and tables when necessary. Output filenames follow a type–difficulty–knowledge-point pattern with a `.png` suffix.

H Annotation Guideline Summary

This summary clarifies the human-labeling policy behind the binary error labels and issue descriptions used throughout the benchmark. Annotators judge a subfigure as erroneous when it contains a visual, mathematical, annotation, notation, or use-risk defect that makes it unsuitable for the assigned

Model	Multiple-choice					Fill-in-the-blank					Free-response				
	Rec.	F1	Spec.	J-Match	J-Partial+	Rec.	F1	Spec.	J-Match	J-Partial+	Rec.	F1	Spec.	J-Match	J-Partial+
Closed-source VLMs															
Claude-Sonnet-4.6	**0.542**	0.426	0.520	**0.086**	0.169	0.529	0.431	0.548	**0.112**	0.161	0.672	0.573	0.520	**0.152**	0.248
Gemini-3.1-Flash-Lite-Preview	0.486	0.439	0.651	0.078	0.176	**0.546**	**0.454**	0.581	0.097	**0.184**	**0.741**	**0.597**	0.471	0.140	**0.265**
GPT-4o	0.363	0.395	0.773	0.035	0.128	0.228	0.278	0.800	0.032	0.087	0.309	0.389	0.802	0.046	0.125
Open-source VLMs															
Qwen2.5-VL-7B	0.050	0.084	0.927	0.000	0.015	0.069	0.112	0.918	0.000	0.015	0.056	0.099	0.945	0.000	0.004
Qwen2.5-VL-32B	0.184	0.250	0.863	0.025	0.065	0.117	0.177	0.901	0.007	0.027	0.169	0.257	0.896	0.019	0.040
Qwen2.5-VL-72B	0.539	**0.472**	0.642	0.058	**0.199**	0.422	0.420	0.713	0.069	0.141	0.484	0.490	0.651	0.104	0.240
LLaVA-OneVision-Qwen2-72B	0.076	0.131	**0.963**	0.000	0.013	0.007	0.014	**0.990**	0.000	0.000	0.010	0.021	**0.996**	0.002	0.006
InternVL3-8B	0.380	0.391	0.728	0.020	0.053	0.223	0.263	0.768	0.005	0.020	0.253	0.319	0.765	0.008	0.035
InternVL3.5-38B	0.169	0.239	0.884	0.025	0.048	0.213	0.271	0.824	0.030	0.057	0.209	0.280	0.798	0.029	0.077
InternVL3-78B	0.086	0.139	0.929	0.005	0.015	0.045	0.079	0.956	0.007	0.012	0.154	0.242	0.914	0.021	0.052

Table 5: Question-format detection and issue-description alignment coverage under the composite-input setting. The best result is shown in bold and the second-best result is underlined for each question format and metric.

Model	Easy					Medium					Hard				
	Rec.	F1	Spec.	J-Match	J-Partial+	Rec.	F1	Spec.	J-Match	J-Partial+	Rec.	F1	Spec.	J-Match	J-Partial+
Closed-source VLMs															
Claude-Sonnet-4.6	0.557	0.455	0.542	**0.119**	0.204	**0.623**	**0.489**	0.528	**0.152**	0.210	0.579	0.495	0.519	0.090	0.178
Gemini-3.1-Flash-Lite-Preview	**0.611**	**0.507**	0.593	0.080	**0.219**	0.580	0.473	0.558	0.123	**0.227**	**0.610**	**0.533**	0.572	**0.115**	0.193
GPT-4o	0.265	0.321	0.803	0.041	0.113	0.319	0.357	0.762	0.048	0.126	0.312	0.388	0.811	0.027	0.105
Open-source VLMs															
Qwen2.5-VL-7B	0.049	0.085	0.950	0.000	0.008	0.068	0.107	0.903	0.000	0.007	0.059	0.101	0.936	0.000	0.017
Qwen2.5-VL-32B	0.183	0.264	0.897	0.026	0.067	0.155	0.218	0.866	0.012	0.034	0.138	0.213	0.897	0.015	0.034
Qwen2.5-VL-72B	0.456	0.453	0.713	0.080	0.206	0.478	0.445	0.660	0.085	0.174	0.505	0.486	0.637	0.073	**0.208**
LLaVA-OneVision-Qwen2-72B	0.028	0.054	**0.993**	0.003	0.010	0.022	0.041	**0.982**	0.000	0.002	0.038	0.070	**0.971**	0.000	0.006
InternVL3-8B	0.286	0.337	0.788	0.013	0.046	0.261	0.293	0.736	0.007	0.029	0.300	0.349	0.736	0.013	0.034
InternVL3.5-38B	0.214	0.285	0.852	0.049	0.090	0.171	0.227	0.828	0.017	0.034	0.208	0.282	0.834	0.021	0.063
InternVL3-78B	0.090	0.152	0.948	0.015	0.023	0.092	0.146	0.918	0.012	0.019	0.111	0.184	0.938	0.008	0.040

Table 6: Difficulty-level detection and issue-description alignment coverage under the composite-input setting. The best result is shown in bold and the second-best result is underlined for each difficulty level and metric.

problem. Acceptable issue descriptions must identify the concrete object or relation, such as a wrong coordinate, an incorrect angle relation, a mislabeled point, a missing required line, or a curve inconsistent with the stated function. Vague comments such as "the figure may be inaccurate" are not sufficient unless they are grounded in a specific visible defect.

I Full Type-Level Diagnostic Results

Figure 5 expands the aggregate type-level analysis in Figure 4. Instead of collapsing over stages, it reports TypeRecall for the most common structural error types within each educational stage, allowing us to inspect which defect categories are detected or missed by each model.

J Type-Level Semantic Alignment

Figure 6 complements the TypeRecall heatmap by measuring whether detected defects are also described in a way that aligns with the human-audited issue descriptions. This appendix figure supports the main-text observation that broad defect coverage does not necessarily imply faithful diagnosis.

K Cropped-Input Ablation Details

Table 9 provides the full cropped-input results corresponding to Figure 3. Each entry reports the cropped-input score followed by the change relative to the composite-image setting, allowing direct comparison of detection and issue-description alignment metrics across models and educational stages.

Role. You are a rigorous expert in mathematical-figure quality inspection.

Input. You are given: (i) the text of a K–12 mathematics problem, (ii) a 2×2 composite figure, and (iii) the position of the subfigure to be inspected. Inspect only the specified subfigure and do not evaluate the other three subfigures. The fixed layout is: subfigure 1 = upper left, subfigure 2 = upper right, subfigure 3 = lower left, and subfigure 4 = lower right.

Inspection criteria. Check whether the specified subfigure contains the mathematical objects required by the problem; whether shapes, points, lines, angles, lengths, coordinates, quantities, and spatial relations are consistent with the problem statement; whether labels such as point names, lengths, angle marks, units, and formulas are correct; whether there are mathematical common-sense errors; whether the image is clear and free of occlusion, overlap, or layout disorder; and whether the figure leaks the answer or provides overly strong solution hints.

Output schema. If no clear issue is found, output:

$$\{\texttt{"has_issue"} : \texttt{false},\ \texttt{"issues"} : []\}.$$

If an issue is found, output:

$$\{\texttt{"has_issue"} : \texttt{true},\ \texttt{"issues"} : [\texttt{"short issue phrase"}]\}.$$

Constraints. Output only one JSON object, without Markdown or additional explanation. The field `issues` must be a string array. If an issue exists, report at most one concise phrase, preferably within 20 Chinese characters in the original prompt, identifying only the problematic location or defect type. If the model cannot judge reliably, return `has_issue=true` and provide a short reason.

Problem text. {*question*}

Table 7: Prompt template used for auxiliary image auditing during dataset construction.

System instruction. You are a strict evaluator for mathematical-figure auditing. Your task is not to re-evaluate the image and not to solve the mathematics problem. Your task is to compare two lists of issue descriptions and determine whether they express the same core defect.

Evaluation principles. Judge semantic equivalence rather than literal word overlap. If one description is more detailed and the other is shorter, assign a high label as long as the core defect is the same. If the prediction captures only part of the core issue, assign a partial label. If the prediction describes a different problem, is directionally wrong, or is unrelated to the human audit, assign a mismatch label. Do not automatically assign mismatch merely because the human audit contains multiple issues and the model predicts only one; if the predicted issue covers a core defect, assign `partial_match` or `match`.

User payload. The Judge receives a JSON object containing `dataset`, `folder`, `topic_key`, `subfigure`, `gold_issues`, and `pred_issues`.

Label definitions.

- `match`: the prediction identifies the same core defect as the human-audited issue, even if phrased differently.
- `partial_match`: the prediction captures part of the defect, misses an important detail, or describes only a related surface symptom.
- `mismatch`: the prediction refers to a different defect, is irrelevant, or contradicts the human-audited issue.

Output schema. Output only a JSON object:

$$\{\texttt{"label"} : \texttt{"match|partial_match|mismatch"},\ \texttt{"score"} : \texttt{0.0–1.0},\ \texttt{"reason"} : \texttt{"brief reason"}\}.$$

Do not output Markdown, code fences, or explanatory prefixes.

Table 8: GPT-4 text-only Judge prompt and labeling criteria used for issue-description alignment evaluation.

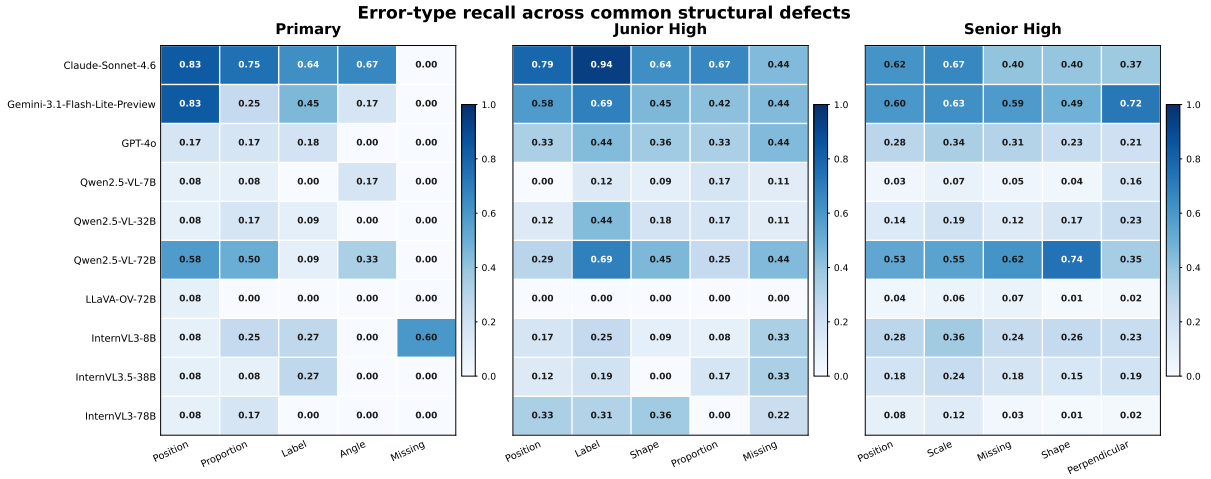


Figure 5: Full type-level detection coverage across common structural error types. Each panel reports TypeRecall for the top-5 non-leakage error types in that educational stage. A cell measures the fraction of human-labeled erroneous subfigures of the given error type that are detected by the model.

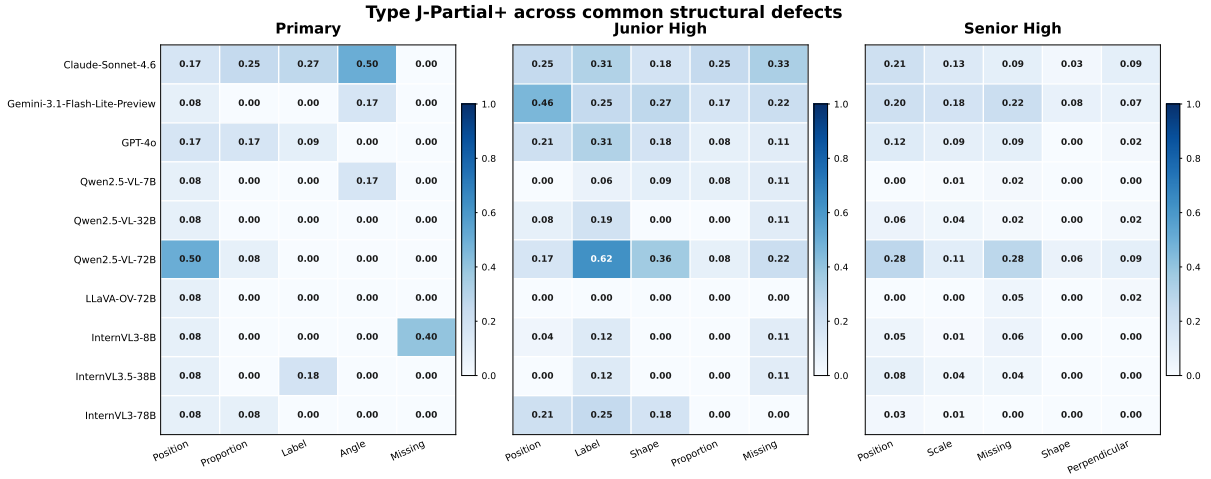


Figure 6: Type-level issue-description alignment coverage across common structural error types. Each cell reports TypeJ-Partial+, computed as the number of MATCH or PARTIALMATCH Judge labels for a given error type divided by all human-labeled erroneous subfigures of that type.

Model	Primary					Junior High					Senior High				
	Rec.	F1	Spec.	J-Match	J-Partial+	Rec.	F1	Spec.	J-Match	J-Partial+	Rec.	F1	Spec.	J-Match	J-Partial+
Closed-source VLMs															
Claude-Sonnet-4.6	**0.816**(+0.193)	0.266(+0.015)	0.703(+0.130)	0.261(+0.091)	0.317(+0.091)	**0.958**(+0.201)	**0.530**(+0.039)	0.569(+0.107)	**0.447**(+0.101)	**0.547**(+0.109)	0.675(+0.126)	0.548(+0.032)	0.436(-0.103)	**0.158**(+0.083)	0.234(+0.083)
Gemini-3.1-Flash-Lite-Preview	0.690(+0.275)	**0.277**(+0.024)	0.489(-0.272)	0.113(+0.094)	0.151(+0.094)	0.886(+0.189)	0.491(-0.063)	0.299(-0.367)	0.438(+0.022)	0.546(+0.022)	**0.875**(+0.281)	**0.603**(+0.071)	0.225(-0.272)	0.131(+0.074)	**0.246**(+0.083)
GPT-4o	0.363(+0.231)	0.170(+0.019)	0.627(-0.293)	0.143(+0.105)	0.199(+0.105)	0.547(+0.206)	0.454(+0.051)	0.615(-0.242)	0.245(+0.121)	0.329(+0.124)	0.465(+0.124)	0.378(-0.038)	0.454(-0.312)	0.070(+0.047)	0.145(+0.047)
Open-source VLMs															
Qwen2.5-VL-7B	0.094(+0.037)	0.124(+0.043)	0.934(-0.023)	0.019(+0.019)	0.019(-0.019)	0.194(+0.097)	0.216(+0.069)	0.754(-0.154)	0.011(+0.011)	0.059(+0.032)	0.419(+0.365)	0.418(+0.323)	0.544(-0.385)	0.011(+0.011)	0.057(+0.051)
Qwen2.5-VL-32B	0.093(-0.001)	0.141(+0.027)	0.963(+0.035)	0.057(+0.057)	0.075(+0.057)	0.336(+0.168)	0.391(+0.146)	0.846(-0.073)	0.124(+0.103)	0.189(+0.146)	0.238(+0.077)	0.333(+0.090)	0.862(-0.005)	0.023(+0.006)	0.077(+0.032)
Qwen2.5-VL-72B	0.203(-0.118)	0.222(-0.011)	0.901(+0.084)	0.094(+0.000)	0.132(+0.000)	0.368(+0.092)	0.387(+0.071)	0.785(-0.024)	0.103(-0.005)	0.184(+0.022)	0.446(-0.081)	0.476(-0.039)	0.667(+0.078)	0.048(-0.025)	0.120(-0.084)
LLaVA-OneVision-Qwen2-72B	0.000(-0.019)	0.000(-0.034)	**0.992**(+0.002)	0.000(+0.000)	0.000(-0.019)	0.048(+0.043)	0.084(+0.073)	0.954(-0.042)	0.000(+0.000)	0.016(+0.016)	0.010(-0.024)	0.021(-0.044)	**0.973**(-0.003)	0.003(+0.002)	0.004(-0.003)
InternVL3-8B	0.578(+0.352)	0.261(+0.058)	0.549(-0.323)	**0.264**(+0.264)	**0.491**(+0.434)	0.400(+0.195)	0.384(+0.118)	0.724(-0.138)	0.016(+0.011)	0.124(+0.097)	0.767(+0.458)	0.576(+0.212)	0.313(-0.379)	0.073(+0.061)	0.205(+0.169)
InternVL3.5-38B	0.015(-0.098)	0.029(-0.106)	0.988(+0.060)	0.019(-0.019)	0.019(-0.057)	0.157(-0.032)	0.244(-0.007)	**0.947**(+0.077)	0.043(+0.032)	0.108(+0.059)	0.145(-0.061)	0.219(-0.064)	0.879(+0.075)	0.012(-0.018)	0.036(-0.027)
InternVL3-78B	0.281(+0.206)	0.273(+0.165)	0.876(-0.083)	0.094(+0.075)	0.151(+0.113)	0.378(+0.135)	0.424(+0.098)	0.835(-0.064)	0.092(+0.065)	0.195(+0.108)	0.478(+0.398)	0.488(+0.349)	0.639(-0.304)	0.054(+0.046)	0.140(+0.123)

Table 9: Full cropped-input results with changes relative to the composite-input setting. Each cell is formatted as $x(y)$, where x is the cropped-input score and y is the change from the composite-input score. Positive values in parentheses indicate improvement after cropping, while negative values indicate degradation.