GradingBench: Evaluating End-to-End Compositional Reasoning of MLLMs for Automated Exam Grading

Yuting Wang[1,2], Zixian Guo[1,2], Weihao You[2], Zhilong Ji[2], Jinfeng Bai[2], and Wangmeng Zuo[1⋆]

[1] Harbin Institute of Technology, Harbin, China
25s003019@stu.hit.edu.cn, wmzuo@hit.edu.cn
[2] Tomorrow Advancing Life, Beijing, China

Abstract. While Multi-modal Large Language Models (MLLMs) have demonstrated superior performance on isolated tasks such as visual question answering, their reliability remains limited in real-world scenarios that require highly compositional visual-language reasoning. A prime example is automated exam grading, a real-world practical task that demands the integration of three core capabilities: detection and localization, text recognition, and reasoning for correctness judgment. Current benchmarks largely focus on isolated tasks and cannot fully evaluate MLLMs' end-to-end ability in such complex settings. To bridge this gap, we present **GradingBench**, a comprehensive benchmark based on automated exam grading in Chinese K–12 education, which systematically evaluates MLLMs across the entire grading pipeline. GradingBench comprises full-page exam papers from real educational settings, containing 3,284 sub-questions annotated with reference answers. We evaluate MLLMs across three levels: single-question, specified-question, and full-page grading. Our experiments show that localization failure is the main bottleneck and reveal limitations in integrated multi-task execution. Although multi-round interaction and supervised fine-tuning bring slight improvements, the limited gains reflect fundamental weaknesses in compositional ability. GradingBench serves as a challenging, application-driven benchmark for the community, underscoring the need for seamless perception-cognition fusion in MLLMs to unlock their potential in practical domains. All code and data are available at `https://github.com/ERRORSEMI/GradingBench`.

Keywords: Multi-modal Large Language Models · Automated Exam Grading · GradingBench

1 Introduction

The development of Multi-modal Large Language Models (MLLMs), such as Gemini-3.1-Pro [13], GPT-5.2 [29], and Qwen3.5 [37], has significantly advanced

⋆ Corresponding author.

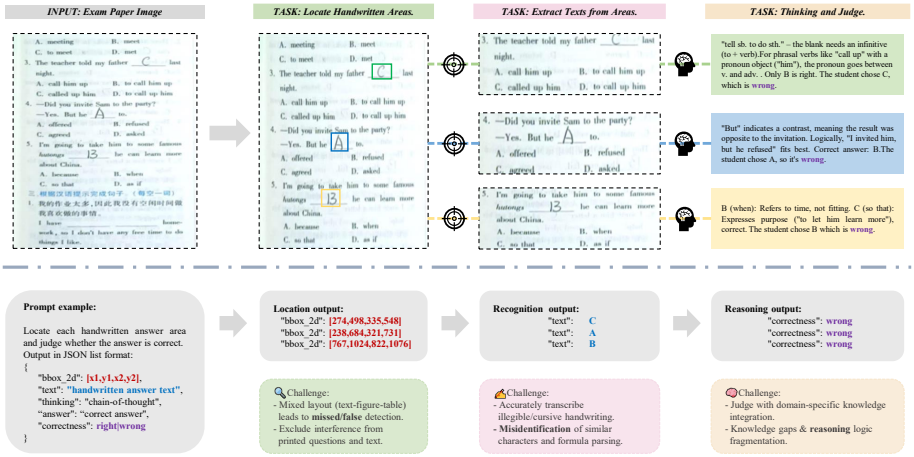

Fig. 1: An overview of the integrated exam grading task. Given a full-page exam image, the model must internally complete the entire workflow: answer region detection and localization, handwritten text recognition, and semantic reasoning for judgment.

the state of the art in a wide range of vision-language tasks, including visual question answering [19–24,31,44] and image captioning [7,25,32]. However, their reliability in complex, real-world tasks requiring the synergistic integration of multiple competencies remains limited.

Most existing vision-language benchmarks, such as TextVQA [31] for text reading or MathVista [19] for mathematical reasoning, evaluate isolated subskills. This decoupled evaluation creates a gap between benchmark performance and a model's proficiency in complex real-world scenarios like exam grading, where perception and cognition are inherently intertwined.

To address this gap, we introduce **GradingBench**, a comprehensive benchmark centered on automated exam grading. As illustrated in Fig. 1, this task requires MLLMs to seamlessly integrate three core capabilities: (1) answer region detection and localization, (2) handwritten text recognition, and (3) semantic reasoning for correctness judgment. Our benchmark requires models to execute this integrated workflow end-to-end in a single forward pass—directly reflecting real-world demands where any component failure impacts the final outcome, posing a formidable challenge to compositional understanding.

GradingBench uses exam papers from scanners and mobile cameras, featuring diverse layouts that reflect realistic conditions. For rigorous evaluation, we provide precise annotations: bounding boxes, handwritten transcriptions, reference answers, and ground-truth grading labels. To systematically probe different facets of perception-cognition integration, we assess models across three progressively complex levels—single-question, specified-question, and full-page grading—under two settings (answer-free and answer-based), forming six evaluation scenarios. Using this framework, we evaluate 18 mainstream MLLMs [1–6,

9, 13, 26, 28, 29, 35–37, 39, 45, 46]. Fig. 2 presents a radar chart comparing six representative models across the six dimensions. Even the powerful Gemini-3.1-Pro achieves only moderate end-to-end accuracy, underscoring the underdeveloped state of multi-competency integration in current MLLMs.

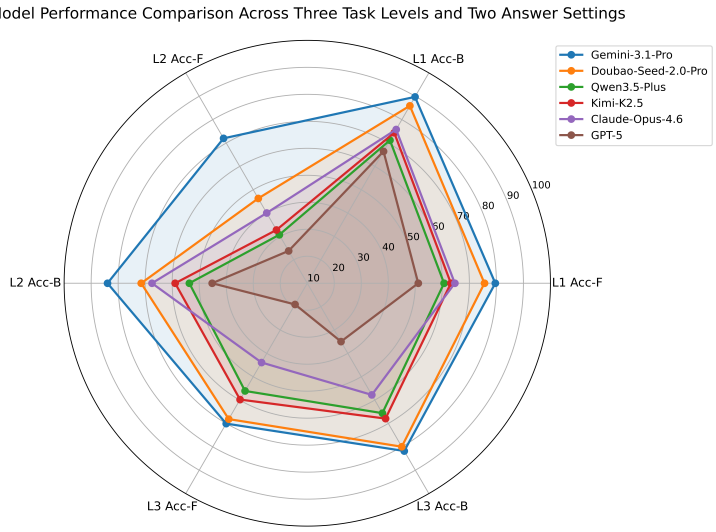

Fig. 2: Radar chart comparing six representative MLLMs across six evaluation scenarios: L1–L3 under both answer-free (F) and answer-based (B) settings. Each axis shows the accuracy for a specific scenario.

The main contributions of this work are as follows:

- We propose GradingBench, the first comprehensive benchmark for end-to-end automated exam grading. It comprises 300 full-page exam images with 3,284 annotated sub-questions, filling a critical gap in evaluating MLLMs on a complex, real-world domain task.
- We perform an extensive evaluation of 18 mainstream MLLMs, diagnosing poor localization and multi-task coordination as key bottlenecks.
- We demonstrate that multi-round interaction and supervised fine-tuning only partially alleviate these limitations, indicating that more fundamental solutions are still required for seamless perception–cognition fusion in real-world grading scenarios.

2 Related Work

2.1 Multi-modal Large Language Models

Recent years have witnessed remarkable progress in Multi-modal Large Language Models (MLLMs), which unify visual encoders and large language backbones to

support joint visual-textual comprehension and logical reasoning. Representative closed-source systems such as Gemini-3.1-Pro [13] and GPT-5.2 [29], alongside open-weight series including Qwen3.5 [37], demonstrate strong performance on generic vision-language tasks covering visual question answering and image captioning. Their evolution integrates stronger perception with sophisticated reasoning, shifting from early focus on feature alignment to handling complex reasoning tasks. However, a critical gap remains as most research emphasizes general-purpose scenarios, overlooking the demands of vertical domains like education and healthcare. These domains require not only generic understanding but also fine-grained visual perception (e.g., recognizing messy handwriting), domain-specific knowledge, and multi-step causal reasoning.

2.2 MLLM Benchmarks

A rich set of vision-language benchmarks evaluates perception and cognitive reasoning separately. For perception tasks, OCRbench [12, 18] targets complex scene text extraction. TextVQA [31] focuses on real-world signage text, and DocVQA [23] covers scanned documents with limited handwritten notes. Moving towards cognition, ChartQA [21] and PlotQA [24] support reasoning over visual charts. Further along, VCR [44], ScienceQA [20] and MathVista [19] test knowledge-heavy multi-step reasoning on simplified visuals.

Despite their utility, these benchmarks often evaluate isolated sub-abilities. Many bypass key perceptual tasks like localization by providing pre-processed inputs, or use oversimplified images lacking real-world complexity. Such designs cannot measure complete pipeline performance, a limitation addressed by GradingBench's end-to-end evaluation paradigm.

2.3 AI for Education

The gap between benchmark capabilities and real-world requirements is particularly evident in education [15, 27]. Recent work has begun to leverage MLLMs for tasks like evaluating handwritten essays [33, 40–42] or mathematical reasoning [8, 14, 16, 19, 30]. However, a recurring limitation persists: these approaches often rely on digitally cropped or pre-transcribed inputs, and avoid the core challenge of parsing full exam sheets with complex layouts and diverse handwriting, failing to reflect real grading complexity. Two closely related educational benchmarks Exams-v [10] and EDU-CIRCUIT-HW [34] only cover limited question types and lack a complete full-page grading pipeline.

Our work integrates these perspectives. GradingBench bridges the identified gaps by requiring models to process raw exam images and execute the complete grading workflow end-to-end—from detection and localization to text recognition, problem solving, and semantic evaluation. This integrated, end-to-end design fills a critical void between isolated task evaluation and real-world application, establishing a more rigorous testbed for developing and assessing MLLMs in practical educational scenarios.

3 The GradingBench Framework

3.1 Task Formulation and Hierarchy

Task Scope. GradingBench supports end-to-end automated grading for tasks with clear correct/incorrect judgments. It includes standard objective questions (e.g., multiple-choice, fill-in-the-blank, true/false) and simple open-ended questions that permit binary correctness evaluation, such as sentence-writing, Chinese reading comprehension, and liberal arts material analysis. We explicitly exclude holistic scoring tasks such as essay grading, which lack standardized automatic evaluation criteria and would complicate the assessment of perceptual and cognitive abilities. By concentrating on binary-judgment tasks, we build a rigorous benchmark and provide a clear foundation for future extensions to more subjective grading.

Dual Evaluation Settings. To mirror a key aspect in real-world educational grading and comprehensively assess a model's adaptability and robustness, we evaluate models under two distinct settings based on the availability of a standard answer key:

- **Answer-Based:** Models are given official reference solutions. This setting evaluates core grading capabilities across four key steps: perceiving, locating, recognizing, and comparing student answers against known correct answers.
- **Answer-Free:** Models must deduce correct solutions independently by leveraging their reasoning capabilities. This evaluates integrated proficiency, testing the grading pipeline alongside domain-specific reasoning.

Task Levels. We further structure our benchmark into three levels of increasing complexity. Each level is evaluated under both the Answer-Based and Answer-Free settings, forming six sub-tasks in total.

- **L1: Single-question Grading** evaluates the model's basic capability in an isolated scenario. The input is a cropped single-question image containing the question and answer region. The output is a JSON list of records, containing answer boxes, recognized text, reasoning chains [38], correct answers, and judgments.
- **L2: Specified-question Grading** evaluates the spatial understanding ability by requiring the model to precisely grade specified questions among irrelevant content on a full page. The input includes a full-page image and the bounding box coordinates of the target question; the output follows the L1 format. This setting reveals a counter-intuitive L2 anomaly (see Sec. 4.1): providing gold coordinates hurts overall accuracy, highlighting fragile spatial-textual coordination.
- **L3: Full-page Grading** evaluates the model's end-to-end automated grading capability. This level most closely resembles real applications, with full-page exam images as input and complete grading results as output.

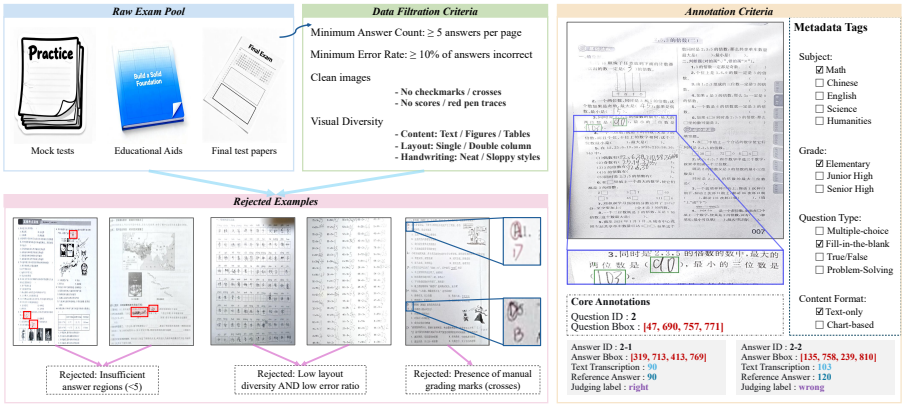

Fig. 3: GradingBench construction pipeline, illustrating the systematic process from raw data collection to structured dataset creation. The core annotation phase encompasses spatial localization, text transcription, reference answer, expert grading, and multi-dimensional metadata tagging.

3.2 Benchmark Construction and Annotation

Data Collection and Selection. We collected full-page exam paper images from real educational settings in China, using both scanning and mobile photography to reflect typical unconstrained conditions. The dataset includes mock tests, educational aids, and final test papers from elementary and secondary schools. Fig. 3 illustrates our benchmark construction pipeline. To ensure data quality, we applied four selection criteria, each addressing a specific requirement for comprehensive evaluation: (1) each paper has at least five handwritten answer regions to ensure sufficient grading content per page; (2) at least 10% of answers per page are incorrect to simulate realistic scenarios, maintain class balance, and ensure the model has enough negative examples to test its discriminative ability; (3) no pre-existing manual grading marks (e.g., ticks, scores) are allowed, as these would interfere with model judgment and introduce noise; (4) diverse layout styles and handwriting variations are included to cover real-world conditions and test model robustness. After filtering, GradingBench contains 300 qualifying full-page images.

Preprocessing. All images underwent distortion correction [43] and adaptive brightness/contrast adjustments to ensure legibility. Unlike benchmarks relying on clean synthetic data, our processing intentionally preserves real-world artifacts—such as uneven lighting and paper creases—to evaluate model robustness under realistic conditions. Personally identifiable information was removed to protect privacy.

Annotation Framework. To support the three-level evaluation tasks defined above, we designed a comprehensive annotation framework with four core components. (Complete annotation guidelines and examples in the Supp.)

Table 1: Overview of GradingBench dataset statistics across the four annotation dimensions. EL: Elementary, JH: Junior High, SH: Senior High, MC: Multiple-choice, FI: Fill-in-the-blank, T/F: True/False, PS: Problem-Solving. Minor discrepancies in totals are due to questions that are difficult to categorize within this dimension.

						Educational Stage			Question Type				Content Format	
Subject	**Total**	✔	✘	✔%	**Images**	**EL**	**JH**	**SH**	**MC**	**FI**	**T/F**	**PS**	**Text**	**Chart**
Science	664	447	217	67%	68	0	490	174	232	383	19	104	328	356
Mathematics	638	444	194	70%	58	581	36	42	108	329	15	158	441	196
Chinese	680	494	186	73%	52	538	78	77	47	514	0	0	615	66
English	647	425	222	66%	53	341	256	53	159	428	31	0	544	83
Humanities	655	422	233	64%	69	0	394	224	295	284	10	50	380	275
Total	3284	2232	1052	68%	300	1460	1254	570	841	1938	75	312	2308	976

- **Localization Metadata:** Bounding boxes in absolute coordinates for precisely delineating each question and answer region ($[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$), used for calculating spatial evaluation metrics.
- **Textual Content:** Two kinds of text are annotated: (1) manual transcriptions of handwritten student answers (including LaTeX for formulas), which provide the ground truth for OCR evaluation; and (2) standard reference answers, which are used for answer-based prompting.
- **Grading Labels:** Expert-annotated binary scores (correct/incorrect) for each answer, providing the definitive judgment for evaluation.
- **Multi-dimensional Metadata:** A structured tagging system across four dimensions—**Subject** (Mathematics, Chinese, English, Science, Humanities), **Educational Stage** (Elementary, Junior High, Senior High), **Question Type** (Multiple-choice, Fill-in-the-blank, True/False, Problem-Solving), and **Content Format** (Text-only, Chart-based)—to enable fine-grained analysis of model capabilities. (More details in the Supp.)

Quality Control. Ten trained annotators with relevant educational backgrounds performed the annotations. Discrepancies between annotators were resolved by senior educational experts, achieving a high inter-annotator Cohen's Kappa of 0.87, which demonstrates reliable labeling quality.
Dataset Statistics. The dataset comprises 300 exam papers with 3,284 subquestions, of which 32.0% are incorrect answers. Tab. 1 provides a detailed breakdown across the four annotation dimensions described above.

3.3 Evaluation Methodology

We design a comprehensive evaluation framework to assess model performance on automated grading. The evaluation framework includes a multi-stage filtering pipeline that sequentially verifies valid spatial correspondence, sufficient text recognition, and correct answer derivation; samples that fail at any stage of the real-world grading scenario are considered incorrect. The final accuracy is calculated according to the end-to-end grading validity. To further visualize the models' fine-grained localization capability, which is important for clear spatial

correspondence between generated grading results and the questions that appear on the exam page, we also calculate a spatial localization metric.

1. Grading Accuracy Calculation.

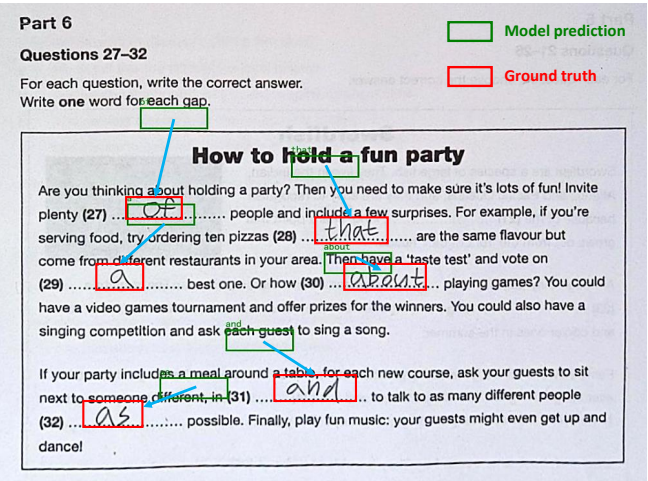

Fig. 4: Center-point matching between model predictions and ground truth. Arrows connect predicted (green) and ground-truth (red) bounding box centers, establishing correspondence while accommodating coordinate offsets.

Center-Point Matching for Semantic Correspondence. To ensure fair assessment of semantic understanding while accounting for precise localization limitations, we employ a center-point matching strategy (Fig. 4). Unlike fixed IoU (which measures spatial overlap of boxes), center-point matching establishes correspondence to evaluate grading correctness. This approach addresses systematic offsets in model predictions: bounding boxes often shift consistently (e.g., upward as in Fig. 4) relative to ground truth, leading to low overlap scores even when the model correctly identifies the semantic region and makes accurate judgments. By using this center-point matching method, our evaluation framework can more effectively handle some minor positioning deviations of the model, thereby objectively evaluating the model's end-to-end marking capability.

Matching is performed greedily from top to bottom: for each ground-truth box, we pair it with the closest predicted box based on Euclidean distance between centers. Our analysis of pilot experiments showed that this simple nearest-neighbor rule successfully pairs over 97% of semantically correct localizations, thus balancing tolerance against spurious matches.

Three-Stage Scoring on the Full Test Set. After establishing correspondence, we compute final grading accuracy using a three-stage rule for each question:

- **Correspondence Check:** If no valid correspondence exists, the question is automatically marked as **incorrect**.
- **Text Recognition Check:** If matched, we calculate the Character Error Rate (CER) between the model-extracted handwritten text and ground-truth text. If CER ≥ 0.7, the question is marked **incorrect**. This threshold was chosen based on an analysis of recognition errors in a development set: CER below this value typically corresponds to recognizable transcriptions, while higher values signify failures.
- **Answer Correctness Check:** If both previous checks pass, verify whether the model correctly derives the answer. In the *answer-free* setting, we use Qwen3-32B [36] for semantic matching between derived and reference answers (validated below). In the *answer-based* setting, the reference is in the prompt, so derivation is trivially correct. If the derived answer does not match the reference (answer-free), the question is automatically marked as **incorrect**. This staged design ensures that a correct judgment is only counted when the model has also correctly derived the answer, thereby avoiding credit for coincidental correct evaluations.
- **Judgment Evaluation:** For questions that survive all three checks, we finally evaluate the model's binary judgment (correct/incorrect) on the student's answer against the ground-truth label. A matching judgment yields a **correct** score; otherwise, it is **incorrect**.

The final Grading Accuracy is then:

$$\text{Acc} = \frac{N_{\text{correct}}}{N_{\text{total}}}$$

where N_{correct} counts questions that passed all three checks and made a correct judgment. By keeping N_{total} constant, this formulation ensures that all models are evaluated on the identical set of questions, and any failure in perception (localization or recognition) or reasoning (incorrect answer derivation) is properly penalized.

Answer Correctness Verification. For answer-free settings, we use Qwen3-32B to automatically check whether the derived answer matches the reference. We validated this approach on 200 stratified samples (covering diverse question types, subjects, and stages) with three expert annotators: agreement with Qwen3-32B reached 98.7%, with only 1.5% variance across models and no systematic bias, confirming its reliability as a proxy for human judgment.

2. Spatial Localization Assessment. To quantify the precision of fine-grained answer region localization within the complete grading pipeline, we compute mean Average Precision averaged over IoU thresholds from 0.2 to 0.5 (mAP@[0.2 : 0.5]), where Average Precision (AP) is calculated separately at each overlap threshold step and then averaged to get a unified localization score:

$$\text{mAP@}[0.2:0.5] = \frac{1}{7}\sum_{k=1}^{7} \text{AP}(0.2 + 0.05(k-1))$$

Table 2: Comprehensive Grading Performance Across All Task Levels. mAP: mean Average Precision at IoU thresholds [0.2, 0.5]. Acc-F is accuracy in answer-free setting, Acc-B is accuracy in answer-based setting, and Overall is the average of all nine metrics (mAP, Acc-F, Acc-B across L1–L3); the **first** and second highest accuracy of models are bolded and underlined, respectively.

	L1 (Single-Question)			L2 (Specified-Question)			L3 (Full-Page)			
Model	**mAP**	**Acc-F**	**Acc-B**	**mAP**	**Acc-F**	**Acc-B**	**mAP**	**Acc-F**	**Acc-B**	**Overall**
Gemini-3.1-Pro	63.03	**79.60**	**89.74**	**47.28**	**72.00**	**83.96**	35.65	**70.06**	**81.78**	**69.23**
Doubao-seed-2.0-Pro	**81.09**	75.58	85.93	30.76	46.30	71.48	**64.15**	68.14	79.95	67.04
Kimi K2.5	20.18	63.15	74.39	16.58	32.81	58.91	25.24	59.77	67.92	46.55
Gemini-2.5-Pro	29.34	67.33	78.11	16.82	38.03	67.08	14.09	50.94	56.09	46.43
Qwen3.5-Plus	22.78	60.63	71.19	14.79	30.73	53.66	36.03	56.09	65.59	45.72
Claude-Opus-4.6	22.30	64.61	75.78	10.40	40.05	67.49	4.59	43.91	57.77	42.99
Doubao-Seed-1.6	10.21	58.50	69.57	10.89	32.34	48.08	15.43	50.79	60.60	39.60
Qwen3-VL-235B-A22B	20.98	51.92	66.96	7.67	17.85	40.53	35.42	47.42	63.68	39.16
GPT-5	20.71	51.07	66.44	8.94	23.84	45.22	2.65	19.07	34.96	30.32
GLM-4.6V	8.55	43.73	61.51	5.78	10.70	26.03	8.23	36.16	43.04	27.08
GPT-5.2	22.53	41.23	63.64	4.21	9.54	28.24	4.00	18.10	33.05	24.95
Claude-Opus-4	12.95	39.66	56.59	4.65	17.48	36.27	1.97	18.88	34.77	24.80
Qwen2.5-VL-72B	6.73	50.37	59.50	1.87	10.96	23.39	1.40	23.33	32.83	23.38
Doubao-1.5-Vision-Pro	10.36	41.14	44.76	4.86	15.90	35.75	4.61	11.18	21.71	21.14
InternVL3-78B	1.78	39.10	49.06	4.08	12.36	22.81	0.87	11.97	19.85	17.99
Gemma-3-27B-IT	0.10	21.68	39.77	3.64	7.46	22.41	0.55	6.91	15.32	13.09
Qwen2.5-VL-7B	0.93	32.67	36.75	1.30	5.18	10.32	0.74	8.38	11.91	12.02
DeepSeek-VL2	2.23	20.07	21.50	0.30	7.22	12.21	0.81	3.32	6.15	8.20

This metric evaluates the overall spatial alignment quality of predicted answer boxes against ground truth, covering multiple overlap criteria rather than a single fixed threshold.

Evaluation Protocol. All models are tested without test-time fine-tuning under unified JSON-format prompts, with bounding box formats adapted per model to eliminate format mismatch bias. Full prompts are given in the Supp. We evaluate six scenarios: L1–L3 with/without reference answers.

4 Experiments

We structure our experimental analysis as a diagnostic-to-remedial progression. First, we benchmark 18 MLLMs' end-to-end grading capability, identify localization failures as the primary bottleneck, and establish a performance hierarchy (Sec. 4.1). Second, we isolate the core capabilities required for automated grading: localization, text recognition, and reasoning, and find that they remain largely intact, implicating multi-task coordination as the underlying issue (Sec. 4.2). Third, we construct a multi-round execution protocol that prompts the model to solve each sub-task one by one. The comparison with single-round end-to-end execution reveals that task decomposition alleviates but does not fully resolve coordination failures (Sec. 4.3). Finally, we explore mitigation through supervised fine-tuning, demonstrating substantial gains while identifying remaining gaps (Sec. 4.4).

4.1 End-to-end Evaluation

We evaluate the 18 MLLMs on 3,284 sub-questions from 300 exam pages using the progressive evaluation framework defined in Sec. 3.3. All models are assessed across six evaluation scenarios (3 task levels $\times$ 2 answer settings) under a strict protocol, with coordinate formats normalized according to each model's specifications. Tab. 2 presents the evaluation results across all task levels. The results reveal significant limitations in MLLMs' comprehensive grading capabilities, with most models achieving low end-to-end accuracy despite possessing individual subtask competencies. (Failure case visualization in the Supp.)

Performance Hierarchy and Localization Bottleneck. Our evaluation reveals a clear performance hierarchy, with Gemini-3.1-Pro achieving the highest overall accuracy (69.23%), followed closely by Doubao-seed-2.0-Pro (67.04%). While these top models demonstrate strong localization, it remains the critical bottleneck for the majority. Across these models, localization failures manifest in two primary modes: **complete misses**, where answers are entirely undetected, and **coordinate offsets**, where positional errors are salvaged by our center-point matching. The prevalence of these failures initiates a performance cascade that severely limits overall grading efficacy.

Task Complexity and Cross-Task Patterns. Performance across task levels reveals nuanced complexity patterns. While performance generally degrades from L1 to L3 in answer-based settings (e.g., Gemini-2.5-Pro: 78.11% $\rightarrow$ 67.08% $\rightarrow$ 56.09%), a notable L2 anomaly occurs in answer-free settings: despite providing explicit coordinates to specify target questions, L2 performance consistently falls below autonomous L3 grading for most models(e.g., Doubao-seed-2.0-Pro: 68.14% L3 vs. 46.30% L2). This anomaly underscores a critical weakness in integrating spatial and textual instructions. The L2 task imposes a unique cognitive burden, as it requires simultaneously parsing textual queries and mapping spatial coordinates—a composite skill that models rarely encounter during pre-training, making it more challenging than autonomous L3 grading.

Reference Answer Utility. The provision of reference answers substantially enhances model performance across the entire pipeline, with consistent improvements observed in final accuracy across all task levels. For Gemini-3.1-Pro, reference answers yield absolute gains of 10.14%, 11.96%, and 11.72% in L1, L2, and L3 respectively. This demonstrates that reference answers not only aid in text recognition but also provide essential contextual knowledge for answer verification, reducing ambiguities in grading. Moreover, reference answers effectively resolve the L2 anomaly. By enabling target identification through robust semantic matching, they bypass the fragile coordinate-parsing step that bottlenecks the answer-free L2 condition. Consequently, the performance gradient normalizes to a monotonic decline (L1 $>$ L2 $>$ L3), confirming that the anomaly was not due to an inherent grading capability ceiling.

In summary, the comprehensive evaluation identifies failed localization (particularly complete misses) as the primary bottleneck. This fundamental limitation in visual grounding outweighs other errors in the grading workflow.

Table 3: Comprehensive sub-capability diagnosis in L1 single-question setting.

Model	Localization mAP	Recognition CER<0.7	Reasoning Accuracy
Gemini-3.1-Pro	**67.72**	92.87	**87.56**
Gemini-2.5-Pro	40.74	90.16	80.48
Qwen3-VL-235B-A22B	41.49	**93.11**	80.60
Doubao-Seed-1.6	13.05	85.54	80.01
Qwen2.5-VL-72B	7.07	84.90	77.38
InternVL3-78B	1.16	79.84	68.14

Table 4: Single-round vs Multi-round on L3 answer-free setting.

Model	Single-round mAP	Single-round Acc-F	Multi-round mAP	Multi-round Acc-F
Gemini-3.1-Pro	**35.65**	**70.06**	39.66	**74.38**
Gemini-2.5-Pro	14.09	50.94	21.31	58.40
Qwen3-VL-235B-A22B	35.42	47.42	**45.49**	58.48
Doubao-Seed-1.6	15.43	50.79	22.97	58.25
Qwen2.5-VL-72B	1.40	23.33	1.34	19.44
InternVL3-78B	0.87	11.97	1.10	14.87

4.2 Capability Isolation & Collaboration Gap

To diagnose whether the localization bottleneck (Sec. 4.1) stems from a perceptual weakness or a coordination failure, we isolate and evaluate the three core capabilities in the L1 setting: localization via answer region detection only; text recognition and reasoning using ground-truth bounding boxes. This controlled setup eliminates cross-task interference, measuring each component's "pure capability ceiling". We employ the same evaluation criteria as in the end-to-end assessment: mAP@[0.2 : 0.5] for localization, a CER threshold of 0.7 for text recognition, and semantic matching for reasoning accuracy.

Isolation Reveals Localization as the Hard Ceiling. As shown in Tab. 3, the performance spread on localization (67.72% vs 1.16%) is 66.56%—the largest gap among all three skills—confirming that visual grounding itself, rather than downstream stages, is the primary constraint. In contrast, text recognition and reasoning exhibit much smaller variance: top models exceed 90% and 87% respectively, with gaps of 13.27% and 19.42%.

Collaboration Failure Analysis. Notably, comparing isolated mAP@[0.2 : 0.5] results (Tab. 3) with full L1 end-to-end performance (Tab. 2) reveals a noticeable performance drop (40.74% $\rightarrow$ 29.34% for Gemini-2.5-Pro), indicating significant multi-task interference beyond intrinsic perceptual limits. We attribute this degradation primarily to the instruction-parsing overhead. The integrated prompt requires the model to process spatial, textual, and semantic information simultaneously, which may overwhelm its capacity and compromise coordinate regression. This suggests that even when individual capabilities are adequate, their coordination in complex workflows remains challenging.

The twofold bottleneck is clear: already-weak standalone localization is further amplified under multi-task load. To examine whether this coordination failure can be mitigated without architectural overhaul, we next decompose the L3 full-page task into three focused, sequential rounds (Sec. 4.3).

4.3 Single-round vs Multi-round Execution

Experimental Design. Building on the identified bottlenecks, we employ a multi-round interaction paradigm to decompose the task and investigate its efficacy in mitigating these issues in realistic settings. We conduct a multi-round interaction evaluation on the L3 full-page grading task, comparing two execution

paradigms: the original **Single-Round** approach requiring simultaneous question detection, localization, recognition, and reasoning versus a **Multi-Round Interaction** paradigm that decomposes the task into three focused stages: pure perception for region localization and transcription, reasoning to determine correct answers, and grading to compare student responses with solutions.

Tab. 4 presents the performance comparison between single-round and multi-round interaction paradigms on the L3 full-page grading task under the answer-free setting.

Moderate but Consistent Performance Gains through Decomposition. The multi-round interaction paradigm demonstrates consistent improvements over single-round execution, with both Gemini-2.5-Pro and Qwen3-VL-235B-A22B achieving comparable final accuracy gains of 7.46% and 11.06%, respectively. This confirms that decomposing the complex grading workflow into distinct phases provides a viable, though limited, pathway to mitigate coordination failures.

Limited Gains and Underlying Challenges. While multi-round interaction provides measurable improvements—evidenced by the increased mAP@[0.2 : 0.5] and final accuracy in Tab. 4—the gains remain modest. The approach confirms that coordination failures are partially malleable through task decomposition, offering a practical workaround for current deployments. However, the limited scale of improvement underscores more fundamental constraints. The persistent performance gap suggests that beyond task decomposition, inherent limitations in perceptual grounding or a lack of relevant compositional reasoning in pre-training data may be the deeper bottlenecks to address in future MLLM development.

4.4 Mitigating Bottlenecks via SFT

To determine if the identified bottlenecks can be mitigated through data-centric intervention, we conduct diagnostic supervised fine-tuning (SFT) [11,17] experiments using Qwen2.5-VL-7B [3] as the base model.

Finetuning Setup. We collected 673 additional exam paper images to construct a diagnostic dataset of approximately 10,000 answer-free examples spanning all three task levels, each including bounding boxes, reference answers, transcribed text, and Chain-of-Thought reasoning chains generated by Gemini-3.1-Pro. (Training examples in the Supp.) The data was organized into three formats: L1 with single-page images, L2 with full-page images and specified bounding boxes, and L3 with full-page images only. We then trained a single model (`Qwen2.5-VL-7B-SFT`) using a unified sequence-to-sequence paradigm on mixed-format data. The fine-tuned model was evaluated on the original test set under the same metrics and protocols as the baseline models.

Tab. 5 presents the comprehensive performance comparison of our unified fine-tuned model against baseline models under the answer-free setting across all three task levels.

The results demonstrate substantial improvements across all task levels. The `Qwen2.5-VL-7B-SFT` model achieves an order-of-magnitude gain in mAP over its

Table 5: Performance comparison of baseline and fine-tuned models across all task levels (answer-free setting)

Model	L1		L2		L3		Average
	mAP	**Acc-F**	**mAP**	**Acc-F**	**mAP**	**Acc-F**	
Gemini-3.1-Pro	<u>63.03</u>	**79.60**	**47.28**	**72.00**	<u>35.65</u>	**70.06**	**61.27**
Gemini-2.5-Pro	29.34	<u>67.33</u>	16.82	<u>38.03</u>	14.09	50.94	36.09
Qwen3-VL-235B-A22B	20.98	51.92	7.67	17.85	35.42	47.42	30.21
Qwen2.5-VL-7B	0.93	32.67	1.30	5.18	0.74	8.38	8.20
Qwen2.5-VL-7B-SFT	**84.81**	62.91	<u>45.72</u>	34.26	**64.49**	<u>52.65</u>	<u>57.47</u>(+49.27)

base version, surpassing Gemini-2.5-Pro by a wide margin and even outperforming Gemini-3.1-Pro in L1 detection (84.81% vs. 63.03%). This confirms that targeted fine-tuning can effectively alleviate perceptual bottlenecks. However, overall accuracy still lags behind Gemini-3.1-Pro (57.47% vs. 61.27%), as gains in semantic reasoning are more modest.

This diagnostic experiment is not intended to compete with state-of-the-art closed-source models, but to demonstrate that even a small open-source model can substantially close the perceptual gap through data-centric intervention. The persistent reasoning gap underscores that while perception can be improved with targeted fine-tuning, deeper compositional reasoning remains a challenge requiring more fundamental advances.

4.5 Discussion and Summary

Our evaluation reveals that while contemporary MLLMs possess strong standalone capabilities, they struggle with integrated perception-cognition in document understanding. Fundamental localization bottlenecks and coordination failures persist in multi-task execution, and although multi-round interaction and supervised fine-tuning offer substantial alleviation, significant gaps remain in complex reasoning and robust visual grounding—core capabilities that require more fundamental solutions.

5 Conclusion

We present GradingBench, a comprehensive benchmark for end-to-end automated exam grading that demands unified perceptual and cognitive reasoning. We test 18 state-of-the-art MLLMs and identify a core limitation: models possess solid standalone abilities but suffer poor localization and cross-task coordination. Though multi-round inference and supervised fine-tuning deliver mild gains, fundamental reasoning deficits persist. GradingBench connects disjoint single-task benchmarks with real grading scenarios, serving as a diagnostic tool to advance unified perception-cognition MLLMs.

The current version of GradingBench focuses on Chinese K-12 education with objective and simple open-ended questions. Beyond benchmarking, we also release a diagnostic SFT dataset of 673 exam paper images. GradingBench's hierarchical framework and staged metrics are language-agnostic and domain-general, readily extensible to other languages and question types. Future directions include extending to fine-grained subjective grading, multilingual expansion, and incorporating interactive grading scenarios.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 2022YFA1004100.

References

1. Anthropic: Introducing Claude 4. https://www.anthropic.com/news/claude-4 (2025), accessed: 2025-11-14
2. Anthropic: Introducing Claude opus 4.6. https://www.anthropic.com/news/claude-opus-4-6 (2026), accessed: 2026-02-27
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. ByteDance: Doubao-1.5-thinking-vision-pro. https://console.volcengine.com/ark/region:ark+cn-beijing/model/detail?Id=doubao-1-5-thinking-vision-pro (2025), volcEngine Model ID: doubao-1-5-thinking-vision-pro-250428, Accessed: 2026-02-27
5. ByteDance: Doubao-seed-1.6. https://console.volcengine.com/ark/region:ark+cn-beijing/model/detail?Id=doubao-seed-1-6 (2025), volcEngine Model ID: doubao-seed-1.6-250615, Accessed: 2026-02-27
6. ByteDance: Doubao-seed-2.0-pro. https://console.volcengine.com/ark/region:ark+cn-beijing/model/detail?Id=doubao-seed-2-0-pro (2026), volcEngine Model ID: doubao-seed-2-0-pro-260215, Accessed: 2026-02-27
7. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. In: European Conference on Computer Vision. pp. 370–387. Springer (2024)
8. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021)
9. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
10. Das, R., Hristov, S., Li, H., Dimitrov, D., Koychev, I., Nakov, P.: Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7768–7791 (2024)

11. Dong, G., Yuan, H., Lu, K., Li, C., Xue, M., Liu, D., Wang, W., Yuan, Z., Zhou, C., Zhou, J.: How abilities in large language models are affected by supervised fine-tuning data composition. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 177–198 (2024)
12. Fu, L., Kuang, Z., Song, J., Huang, M., Yang, B., Li, Y., Zhu, L., Luo, Q., Wang, X., Lu, H., et al.: Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. arXiv preprint arXiv:2501.00321 (2024)
13. Google DeepMind: Gemini 3.1 pro. `https://deepmind.google/models/gemini/pro/` (2026), accessed: 2026-02-27
14. Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., Steinhardt, J.: Measuring mathematical problem solving with the math dataset. arXiv preprint arXiv:2103.03874 (2021)
15. Kim, J., Lee, H., Cho, Y.H.: Learning design to support student-ai collaboration: Perspectives of leading teachers for ai in education. Education and information technologies **27**(5), 6069–6104 (2022)
16. Li, J., Beeching, E., Tunstall, L., Lipkin, B., Soletskyi, R., Huang, S., Rasul, K., Yu, L., Jiang, A.Q., Shen, Z., et al.: Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. Hugging Face repository **13**(9), 9 (2024)
17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
18. Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.C., Liu, C.L., Jin, L., Bai, X.: Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences **67**(12), 220102 (2024)
19. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
20. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems **35**, 2507–2521 (2022)
21. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244 (2022)
22. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022)
23. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021)
24. Methani, N., Ganguly, P., Khapra, M.M., Kumar, P.: Plotqa: Reasoning over scientific plots. In: Proceedings of the ieee/cvf winter conference on applications of computer vision. pp. 1527–1536 (2020)
25. Mokady, R., Hertz, A., Bermano, A.H.: Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734 (2021)
26. Moonshot AI: Kimi k2.5: Ai that sees, codes, and works like an expert. `https://www.kimi.com/ai-models/kimi-k2-5` (2026), accessed: 2026-02-27

27. Ng, D.T.K., Lee, M., Tan, R.J.Y., Hu, X., Downie, J.S., Chu, S.K.W.: A review of ai teaching and learning from 2000 to 2020. Education and Information Technologies **28**(7), 8445–8501 (2023)
28. OpenAI: Introducing GPT-5. https://openai.com/index/introducing-gpt-5/ (2025), accessed: 2026-02-27
29. OpenAI: Introducing GPT-5.2. https://openai.com/index/introducing-gpt-5-2/ (2025), accessed: 2026-02-27
30. Qiao, R., Tan, Q., Yang, P., Wang, Y., Wang, X., Wan, E., Zhou, S., Dong, G., Zeng, Y., Xu, Y., et al.: We-math 2.0: A versatile mathbook system for incentivizing visual mathematical reasoning. arXiv preprint arXiv:2508.10433 (2025)
31. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
32. Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., Cucchiara, R.: From show to tell: A survey on deep learning-based image captioning. IEEE transactions on pattern analysis and machine intelligence **45**(1), 539–559 (2022)
33. Su, J., Yan, Y., Fu, F., Zhang, H., Ye, J., Liu, X., Huo, J., Zhou, H., Hu, X.: Essayjudge: A multi-granular benchmark for assessing automated essay scoring capabilities of multimodal large language models. arXiv preprint arXiv:2502.11916 (2025)
34. Sun, W., Chen, L., Cai, Y., Xie, H., Zeng, Y., Zhang, Y.: Edu-circuit-hw: Evaluating multimodal large language models on real-world university-level stem student handwritten solutions. arXiv preprint arXiv:2602.00095 (2026)
35. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
36. Team, Q.: Qwen3-vl: Sharper vision, deeper thought, broader action. https://qwen.ai/blog?id=qwen3-vl (2025), accessed: 2026-02-27
37. Team, Q.: Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5 (2026), accessed: 2026-02-27
38. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
39. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)
40. Xiao, C., Ma, W., Xu, S.X., Zhang, K., Wang, Y., Fu, Q.: From automation to augmentation: Large language models elevating essay scoring landscape. CoRR (2024)
41. Yang, K., Raković, M., Li, Y., Guan, Q., Gašević, D., Chen, G.: Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability. In: Proceedings of the aaai conference on artificial intelligence. vol. 38, pp. 22466–22474 (2024)
42. Ye, J., Wang, S., Zou, D., Yan, Y., Wang, K., Zheng, H.T., Xu, Z., King, I., Yu, P.S., Wen, Q.: Position: Llms can be good tutors in foreign language education. arXiv preprint arXiv:2502.05467 (2025)
43. Yu, F., Xie, Y., Wu, L., Wen, Y., Wang, G., Ren, S., Chen, X., Mao, J., Li, W.: Docreal: Robust document dewarping of real-life images via attention-enhanced control point prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 665–674 (2024)

44. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6720–6731 (2019)
45. Zhipu AI: Glm-4.6v: Visual understanding model. `https://docs.bigmodel.cn/cn/guide/models/vlm/glm-4.6v` (2026), accessed: 2026-02-28
46. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)