EduMARS: Can Vision-Language Models Grade Like Teachers? Benchmarking Multimodal, Rubric-Based Assessment on Chinese K-12 Answers

Xuan Zhao*, **Jiashun Chen***, **Wanting Xu**, **Huiyuan Yan**, **Chaowei Fang**, **Xing Wei**†

Xi'an Jiaotong University

Correspondence: weixing@mail.xjtu.edu.cn

Abstract

Automated grading of student work is a critical application of AI in education. However, existing benchmarks fall short in evaluating models on **realistic**, **cognitively demanding** tasks. Most rely on synthetic, well-structured text inputs, overlooking the multimodal, error-prone, and often handwritten nature of real student responses, especially in K-12 settings. We introduce **EduMARS**, a multimodal benchmark designed for rubric-aligned evaluation of real Chinese K-12 student answers. The dataset contains over 4,500 authentic responses from high-stakes exams across eight subjects, featuring **noisy handwriting**, **mixed-layout diagrams**, **mathematical expressions**, and **narrative reasoning**. Each response is meticulously annotated by expert teachers using step-wise scoring rubrics, error classifications, and keypoint mappings, providing fine-grained supervision aligned with real-world pedagogical practices. We evaluated existing SOTA MLLMs across the dimensions of final score and the reasoning process of grading, reveals a significant gap between existing SOTA MLLMs and human-level performance. To bridge this performance gap, we propose the Retrieval-Augmented Adaptive-Rubric Grading (RARG), enabling models to emulate expert grading logic by dynamically synthesizing case-specific evaluation schemas. RARG effectively enhances the performance and interpretability of various MLLMs on EduMARS, surpassing in-context learning and chain-of-thought.

1 Introduction

Automated grading of student answers is a long-standing and increasingly high-stakes problem in education. Recent advances in multimodal large language models (MLLMs) have sparked growing interest in applying such models to educational assessment, particularly for open-ended and cognitively demanding tasks in K–12 settings. Automating the evaluation of free-form responses promises improved efficiency and consistency compared to manual grading, especially at scale.

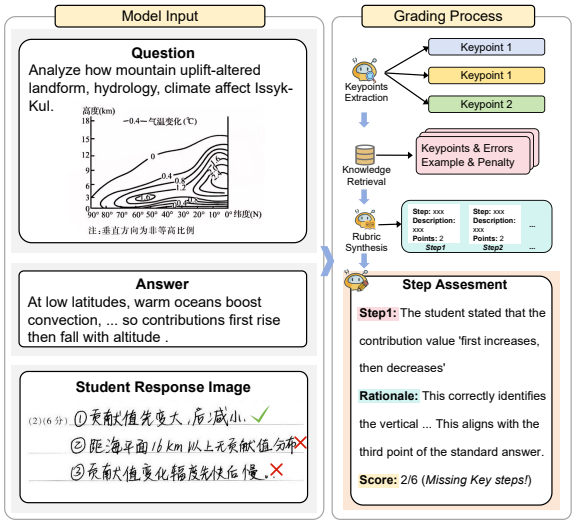

Figure 1: Overview of the real-world exam benchmark and the Retrieval-Augmented Adaptive-Rubric Grading (RARG) pipeline.

However, automated grading involves more than checking final correctness. In real-world K–12 examinations, teachers assess student responses by examining intermediate reasoning steps, assigning partial credit, and identifying conceptual errors according to detailed scoring rubrics. This rubric-aligned grading process reflects pedagogical judgment and step-wise evaluation, rather than holistic outcome estimation. While MLLMs have demonstrated strong multimodal reasoning capabilities, it remains unclear whether they can faithfully replicate such fine-grained, teacher-like grading processes.

Existing benchmarks for automated grading fall short in capturing this setting along two key dimensions. First, many benchmarks (Mathias and

* Equal contribution.
† Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 9561–9583
July 2-7, 2026

Benchmark	Modality	Source	Step Annot.	Annot. Type	Subjects	Step Content
CriticBench (Lin et al., 2024a)	Text	Synthetic	✗	N/A	N/A	N/A
MATHCHECK-GSM (Zhou et al., 2024)	Text	Synthetic	✓	Synthetic	Single	First error positions
PROCESSBENCH (Zheng et al., 2025)	Text	Synthetic	✓	Human	Single	Errors
DrawEduMath (Baral et al., 2024)	Image	Human	✗	Human	Single	N/A
SAS-Bench (Lai et al., 2025a)	Text	Synthetic	✓	Human	Multi	Predefined errors
EduMARS (Ours)	**Image**	**Human**	✓	**Human**	**Multi**	**Scoring rationales**

Table 1: Comparison of EduMARS with representative benchmarks.

Bhattacharyya, 2018; Cobbe et al., 2021; Lu et al., 2022) rely on synthetic or sanitized text inputs, overlooking the visual and structural complexity of authentic student responses, such as handwritten solutions, mixed layouts, and diagram-based reasoning. Furthermore, even when visual data is incorporated, existing datasets (Lu et al., 2024; Li et al., 2025) often suffer from limited subject coverage, failing to encompass the diverse reasoning patterns inherent in various academic disciplines. Second, prior datasets (Wang et al., 2024c; Zhu et al., 2024; Mirzadeh et al., 2025) often depend on model-generated annotations or automated labeling, which can bias evaluation and fail to accommodate alternative valid solution strategies. Moreover, even when human annotations are available, they are typically not tailored for educational grading, lacking the granular, step-by-step rationales required for professional assessment (Lightman et al., 2023; Wang et al., 2024b; Ke and Ng, 2019). Even SAS-Bench (Lai et al., 2025a), which provides step-level annotations for educational grading, offers only predefined error categories rather than the actual scoring rationales used by teachers during grading. As a result, current benchmarks inadequately measure a model's ability to follow rubric-aligned grading procedures in realistic educational scenarios.

To address these limitations, we introduce **EduMARS** (**Edu**cation-oriented **M**ultimodal **A**ssessment with **R**ubric-based **S**coring), a large-scale multimodal benchmark designed for rubric-aligned assessment of authentic K–12 student responses. EduMARS consists of over 4,500 real-world exam submissions collected from high-stakes examinations across eight subjects. Each response is annotated by expert teachers with step-wise scoring rubrics, partial credit assignments, and grading rationales, enabling direct evaluation of both final scores and the underlying grading process. This structure allows EduMARS to assess whether models adhere to human grading standards beyond surface-level correctness.

We conduct a comprehensive evaluation of state-of-the-art multimodal large language models (MLLMs) on EduMARS, covering both final-score prediction and the reconstruction of human grading rationales. The results reveal a significant performance gap between these models and expert human graders, especially in tasks involving partial credit allocation, subtle misconception diagnosis, and holistic multimodal interpretation.

To this end, we propose **RARG** (**R**etrieval-**A**ugmented Adaptive-**R**ubric **G**rading), mimicking expert workflows to extract concepts, retrieve knowledge, and synthesize case-specific rubrics for evidence-grounded grading. This significantly improves human alignment, outperforming standard in-context learning and multi-turn chain-of-thought baselines. (Brown et al., 2020; Wei et al., 2022; Wang et al., 2023).

Our primary contributions are summarized as follows:

- We present **EduMARS**, a large-scale multimodal benchmark featuring over 4,500 authentic, rubric-annotated K-12 exam responses across eight major subjects, setting a new standard for realistic educational evaluation.
- Extensive evaluations on EduMARS reveal a significant gap between MLLMs and humans in automated grading, and highlights the notable degradation in the grading process compared to the final score.
- We propose **RARG** (Retrieval-Augmented Adaptive-Rubric Grading), which dynamically synthesizes rubrics via retrieval to effectively improve the alignment with human grading standards.

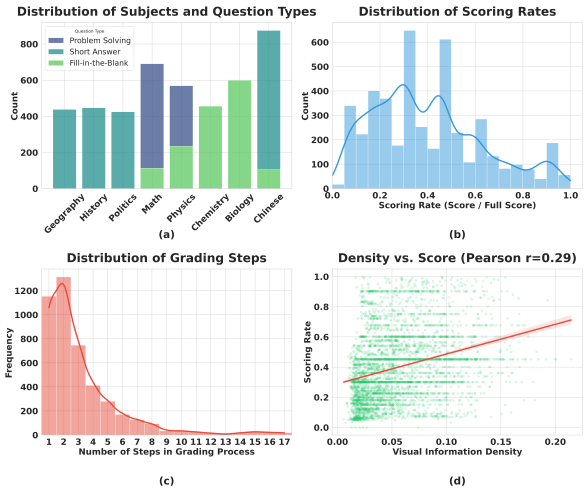

Figure 2: Statistics of the EduMARS

2 Related Works

2.1 LLM-as-a-judge

The LLM-as-a-judge paradigm has shifted from semantic similarity metrics like BERTScore (Zhang et al., 2020) to generative adjudication frameworks such as GPTScore (Fu et al., 2024) and G-Eval (Liu et al., 2023). Recent advancements focus on fine-grained instruction-following via Prometheus (Kim et al., 2024) and dynamic benchmarks like JudgeBench (Tan et al., 2025). In educational contexts, research is moving toward Cognitive Process Tracing, though LLMs still face challenges with dimension-dependency (Shen et al., 2023) and self-preference bias (Wataoka et al., 2025). While these works primarily evaluate models on synthetic or well-structured text , our work distinguishes itself by introducing EduMARS, which benchmarks multimodal, rubric-aligned grading on authentic, handwritten student responses .Although frameworks like FLASK (Ye et al., 2024) target broad skill-set alignment, they often lack the explicit, step-wise rigor necessitated by professional pedagogical standards. Our proposed RARG framework addresses this limitation by anchoring model reasoning to verified educational rubrics, thereby ensuring both process fidelity and scoring precision.

2.2 Benchmark of Automated Grading

From Textual Alignment to Multimodal Logic Auditing. Grading benchmarks are evolving from text-centric platforms like CriticBench (Lin et al., 2024b) to complex multimodal adjudication. Key developments include MathVista (Lu et al., 2024) for visual mathematical reasoning , and MathVerse (Zhang et al., 2024), which investigates whether MLLMs "truly see" diagrams. Recent efforts have reached Olympiad-grade challenges with MATH-Vision (Wang et al., 2024a) and Omni-MATH (Gao et al., 2025). To ensure reliability, ProcessBench (Zheng et al., 2025) identifies reasoning errors , while SAS-Bench (Lai et al., 2025b) offers fine-grained short answer scoring. Additionally, Conformal Prediction (Sheng et al., 2025) provides a framework for analyzing evaluation uncertainty. Unlike benchmarks relying on synthetic data or holistic scoring, EduMARS leverages 4,500+ authentic handwritten K-12 responses and introduces step-wise, rubric-aligned supervision to bridge the gap between model estimation and expert judgment.

3 EduMARS

3.1 Data Collection

We collected over 2,000 response sheets from more than 100 examinees across two distinct mock examinations for the national college entrance examination. To ensure privacy compliance, all personally identifiable information, such as candidate names and ID numbers, was removed. Subsequently, the sheets were digitized using professional scanners. We then segmented the full-page images into individual question-response images based on predefined answer zones.

3.2 Annotation Pipeline

To ensure the pedagogical rigor of our dataset, we recruited 10 expert annotators with extensive teaching experience to conduct a two-phase annotation process. Initially, for each question, experts formulated a comprehensive scoring rubric detailing the required key points and credit distribution. Subsequently, guided by these rubrics, the annotators performed a careful evaluation of each student response. Unlike traditional scoring, which provides only a final grade, our experts annotated three distinct components: (1) key response steps corresponding to the scoring rubric, (2) step-wise scores for each key step,and (3) step-specific grading rationales providing justifications for each partial score. Finally, the overall score was derived by aggregating these individual step components

Notations Each assessment sample $\mathcal{D}_i$ comprises a student response image I_i, a question context Q_i (containing the problem statement and standard reference answer), and a comprehensive expert anno-

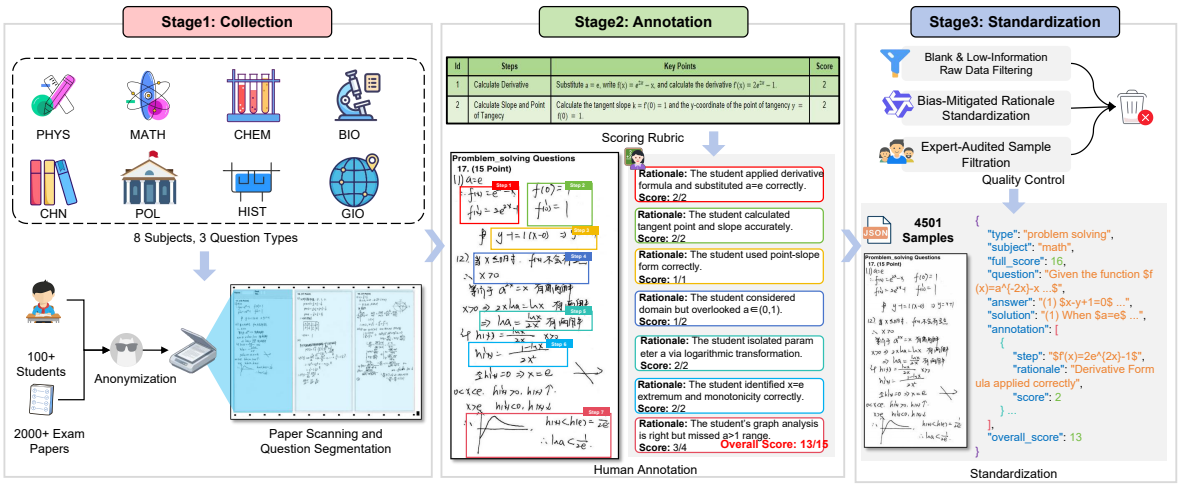

Figure 3: The pipeline illustrates the transition from raw physical exam papers to standardized JSON-formatted assessment samples. Key phases include large-scale multi-subject collection, human-centric step-by-step scoring, and an automated-expert hybrid audit for rationale standardization.

tation $\mathcal{A}_i$. This annotation is formalized as a tuple $\mathcal{A}_i = \langle S_{final}, \mathcal{K}, \mathcal{S}_{step}, \mathcal{E} \rangle$, where $S_{final} \in \mathbb{R}$ denotes the final aggregated score. The sets $\mathcal{K}$, $\mathcal{S}_{step}$, and $\mathcal{E}$ represent the key response steps, step-wise scores, and step-wise grading rationales, respectively. Crucially, the elements across these three sets are strictly aligned, establishing a one-to-one correspondence between each key segment, score, and rationale.

3.3 Quality Control And Standardization

We enforced strict quality control by first filtering invalid raw data and standardizing rationales via Qwen3-32B-Instruct (Yang et al., 2025) to mitigate bias. Subsequently, 10 experts conducted a manual audit to verify logical consistency, yielding a final dataset of 4,501 high-quality samples.

3.4 Benchmark Comparison

Table 1 contrasts EduMARS with existing educational evaluation benchmarks across five key dimensions, including data modality, grading granularity, reasoning type, and expert validation. In contrast to many prior datasets that rely on synthetic examples or text-only inputs, EduMARS distinguishes itself by uniquely integrating real-world handwritten and printed image inputs with fine-grained, expert-verified, multi-disciplinary reasoning chains that closely align with authentic human grading practices.

3.5 Dataset Statistics

Figure 2 presents the key statistics of EduMARS. The dataset includes 4,501 authentic student responses across 8 subjects. Scoring rates average 38.7%, exhibiting a distribution that closely mirrors real-world examination scenarios. The rationale steps average 2.8 per response, ranging from 1 to 17 steps—reflecting the complexity of human grading.. This wide range captures diverse grading processes and complex reasoning chains, underscoring the critical need for models to effectively discern valid reasoning steps amidst visual noise.

3.6 Evaluation

Pipeline We input the student's original handwritten image I I directly into the model, using a prompt that contains only the question and the standard answer. The model then analyzes the handwriting to generate its own scoring rationale. To evaluate its performance comprehensively, we compare the model's output against expert-verified labels along two key dimensions: final score accuracy and fidelity in replicating the human grading process—reflecting not just what score to assign, but how that judgment is reached.

Model Output Definition The model output is structured to align directly with the expert annotation format $\mathcal{A}$. We formalize the prediction as a tuple $\hat{\mathcal{A}} = \langle \hat{S}_{\text{final}}, \hat{\mathcal{K}}, \hat{\mathcal{S}}_{\text{step}}, \hat{\mathcal{E}} \rangle$, where $\hat{S}_{\text{final}} \in \mathbb{R}$ represents the predicted final score. The sets $\hat{\mathcal{K}}$, $\hat{\mathcal{S}}_{\text{step}}$, and $\hat{\mathcal{E}}$ denote the predicted key response segments, step-wise scores, and generated scoring ra-

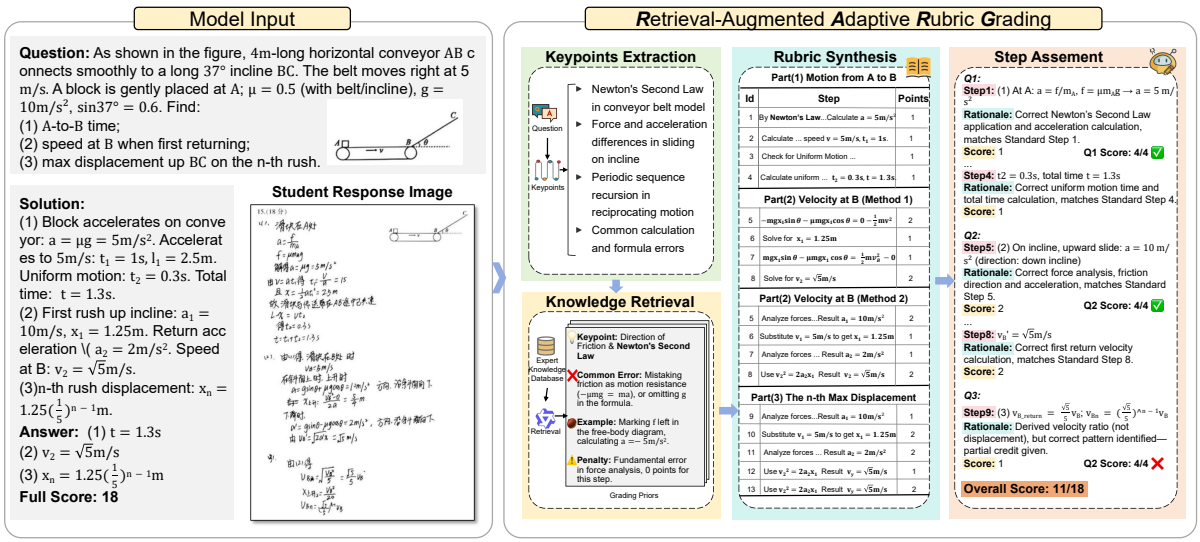

Figure 4: Architecture of the RARG framework. RARG extracts keypoints to retrieve grading priors, synthesizes an adaptive rubric, and evaluates student responses via step-wise verification.

tionales, respectively. Similar to the ground truth, a strict one-to-one correspondence is maintained across the elements of $\hat{\mathcal{K}}$, $\hat{\mathcal{S}}_{\text{step}}$, and $\hat{\mathcal{E}}$.

Final Score Evaluation Metrics To evaluate Score Evaluation performance, we employ three complementary metrics.

First, Spearman's Rank Correlation (ρ) assesses monotonic consistency:

$$\text{Spearman} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}, \tag{1}$$

where N is the total sample size, and $d_i = \text{rank}(S_{\text{final},i}) - \text{rank}(\hat{S}_{\text{final},i})$ denotes the rank difference between the ground truth and predicted scores for the i-th sample.

Second, Normalized Mean Absolute Error (NMAE) measures absolute deviation relative to question difficulty:

$$\text{NMAE} = \frac{1}{N}\sum_{i=1}^{N} \frac{|S_{\text{final},i} - \hat{S}_{\text{final},i}|}{S_{\text{max},i}}, \tag{2}$$

where $S_{\text{max},i}$ is the maximum possible score for the i-th question, normalizing the absolute error between the ground truth $S_{\text{final},i}$ and prediction $\hat{S}_{\text{final},i}$.

Finally, to handle class imbalance, we utilize the Weighted F1:

$$\text{Weighted F1} = \sum_{k=1}^{|\mathcal{C}|} \frac{N_k}{N} \cdot \frac{2P_kR_k}{P_k + R_k}, \tag{3}$$

where $\mathcal{C}$ represents the set of unique score values in the ground truth, treating each distinct score as a separate class. Here, n_k is the count of samples with score k, serving as the class weight, while P_k and R_k represent the precision and recall for score class k. This formulation allows us to account for label imbalance and compute a reliable class-aware summary of performance.

Grading Process Evaluation Metrics To evaluate the scoring process, we represent predictions and ground truth as sets of reasoning units, denoted as $\hat{\mathcal{U}} = \{\hat{u}_1, \ldots, \hat{u}_{\hat{n}}\}$ and $\mathcal{U} = \{u_1, \ldots, u_n\}$, respectively. Each unit $u_i = \langle k_i, e_i, s_i \rangle$ consists of a key step, rationale, and score. The set of aligned pairs $\mathcal{M} \subseteq \hat{\mathcal{U}} \times \mathcal{U}$ is determined via a three-stage verification process: a predicted unit $\hat{u}_i$ aligns with u_j if (1) the key steps match (exact string alignment for STEM or $\text{Sim}(\hat{k}_i, k_j) > \tau_{\text{key}}$ for humanities); (2) the rationales satisfy $\text{Sim}(\hat{e}_i, e_j) > \tau_{\text{rat}}$; and (3) the scores are identical ($\hat{s}_i = s_j$). We enforce a strict one-to-one mapping constraint, ensuring that each predicted unit matches at most one ground truth unit. Semantic similarities are computed using Qwen3-embedding-4B. By aggregating aligned pairs across all test samples, we report three global metrics. The Jaccard index penalizes hallucinations:

$$\text{Jaccard} = \frac{1}{N}\sum_{i=1}^{N} \frac{|\mathcal{M}_i|}{|\hat{\mathcal{U}}_i| + |\mathcal{U}_i| - |\mathcal{M}_i|} \tag{4}$$

$$\text{Recall} = \frac{1}{N}\sum_{i=1}^{N} \frac{|\mathcal{M}_i|}{|\mathcal{U}_i|} \tag{5}$$

Method	Final Score Evaluation			Grading Process Evaluation		
	Spearman↑	NMAE↓	WeightedF1↑	Jaccard↑	Recall↑	F1↑
Human Level	0.945	0.092	0.824	0.892	0.965	0.923
Open-source Models						
Qwen3-VL-4B-Instruct	0.419	0.308	0.154	0.134	0.179	0.140
Qwen3-VL-8B-Instruct	0.479	0.291	0.200	0.161	0.270	0.213
InternVL3.5-14B	0.206	0.367	0.122	0.034	0.054	0.045
DeepSeek-VL2	0.454	0.292	0.179	0.129	0.235	0.228
Qwen3-VL-32B-Instruct	0.532	0.250	0.255	0.192	0.332	0.253
InternVL3.5-38B	0.276	0.325	0.145	0.069	0.112	0.129
GLM-4.6V	0.535	0.248	0.260	0.195	0.335	0.255
Qwen3-VL-235B-Instruct	0.590	0.223	0.269	0.247	0.328	0.251
Closed-source Models						
GPT-5†	0.638	**0.191**	0.316	0.272	0.365	0.338
Gemini-3-Flash-Preview†	**0.645**	0.204	**0.325**	**0.294**	**0.375**	**0.356**

Table 2: Evaluation across Final Score and Grading Process Performance. Most models are evaluated on the full test set; those marked with † are evaluated on a subset of 1,200 samples due to API constraints. The best performance is highlighted in **bold**.

$$\text{F1} = \frac{1}{N}\sum_{i=1}^{N}\frac{2\cdot|\mathcal{M}_i|}{|\hat{\mathcal{U}}_i|+|\mathcal{U}_i|} \tag{6}$$

4 RARG

Since existing methods relying on in-context exemplar injection often disproportionately prioritize outcome metrics while neglecting the explicit modeling of the intermediate reasoning process, we propose the Retrieval-Augmented Adaptive-Rubric Grading (RARG) framework to bridge this gap by transitioning from shallow pattern matching to a structured assessment grounded in dynamically synthesized, expert-aligned verification standards.

4.1 Overview of RARG

As illustrated in Figure 4, RARG aligns grading with expert cognition through four sequential modules. The pipeline begins with keypoints extraction and knowledge retrieval to fetch granular error patterns. These insights drive rubric synthesis to generate a case-specific scoring schema, enabling the final step assessment to evaluate responses with detailed rationales.

4.2 Pipeline

Keypoint Extraction The framework initiates by taking the question and standard solution as input. It decomposes these into semantic keypoints, which represent the core structural concepts and constraints of the problem. This structured representation ensures consistent and interpretable modeling, thereby enabling robust downstream alignment, evaluation, and fine-grained feedback generation.

Knowledge Retrieval Using the extracted Keypoints as queries, the system searches the expert knowledge database to retrieve grading priors incorporating verified error patterns and penalty protocols—thereby grounding the subsequent generation in pedagogical expertise.

Rubric Synthesis Conditioned on the question, standard solution, and the retrieved grading priors, the model synthesizes an adaptive Rubric. This structured schema transforms abstract logic into a case-specific, step-wise scoring table.

Step Assessment Finally, the pipeline takes the student response and the synthesized adaptive rubric as input. It performs a step-wise comparison to output the final grading result.

Model	Input Modality	Method	Final Score Evaluation			Grading Process Evaluation		
			Spear.↑	NMAE↓	W-F1↑	Jaccard↑	Recall↑	F1↑
Qwen3-VL-235B-Instruct	Raw Image	Zero-Shot	0.590	0.223	0.269	0.247	0.328	0.251
	Image + OCR	Zero-Shot	0.605	0.215	0.282	0.260	0.345	0.265
	Raw Image	3-Shots	0.618	0.208	0.295	0.275	0.362	0.282
	Raw Image	MT-CoT	0.635	0.198	0.310	0.292	0.380	0.301
	Raw Image	RARG	**0.755**	**0.112**	**0.452**	**0.435**	**0.582**	**0.472**
GPT-5†	Raw Image	Zero-Shot	0.638	0.191	0.316	0.272	0.365	0.338
	Image + OCR	Zero-Shot	0.652	0.185	0.332	0.285	0.380	0.345
	Raw Image	3-Shots	0.665	0.178	0.350	0.302	0.405	0.362
	Raw Image	MT-CoT	0.680	0.170	0.372	0.320	0.428	0.385
	Raw Image	RARG	**0.825**	**0.082**	**0.548**	**0.512**	**0.695**	**0.584**
Gemini-3-Flash-Preview†	Raw Image	Zero-Shot	0.645	0.204	0.325	0.294	0.375	0.356
	Image + OCR	Zero-shot	0.658	0.198	0.340	0.310	0.395	0.368
	Raw Image	3-Shots	0.672	0.185	0.355	0.325	0.412	0.385
	Raw Image	MT-CoT	0.705	0.165	0.392	0.358	0.448	0.415
	Raw Image	RARG	**0.842**	**0.088**	**0.565**	**0.538**	**0.720**	**0.612**

Table 3: We take zero-shot inference with raw images as the baseline. Under this zero-shot setting, we assess the impact of OCR augmentation; using raw images, we further compare our RARG framework against strong prompting baselines. The best result in each column is in **bold**; models marked with † are evaluated on 1,200 samples due to API constraints.

This structured decomposition effectively shifts the paradigm from implicit holistic estimation to precise, step-wise verification.

5 Experiments

5.1 Baseline

Evaluated MLLMs We conduct a comprehensive evaluation on a diverse set of Large Vision Language Models , comprising seven open-source models and two closed-source models. The open-source lineup includes the Qwen3-VL-Instruct series (4B, 8B, 32B, and 235B) (Bai et al., 2025), the InternVL3.5 series (14B and 38B) (Wang et al., 2025), and DeepSeek-VL2 (Wu et al., 2024). For closed-source models, we include GPT-5 (OpenAI, 2025) and Gemini3-Flash (Gemini Team, 2025). We explicitly set the temperature parameter to 0.1 for all model evaluations.

Human Baseline To establish a rigorous benchmark for human-level performance on this grading task, we introduced a human baseline. Specifically, we recruited five professional teachers to independently evaluate the student responses.To ensure a fair comparison, these experts were provided with the exact same input context as the models. Their aggregated performance serves as a high-quality reference point (Human Level) to assess the gap between current state-of-the-art MLLMs and expert human graders.For the metric calculations, we set the similarity thresholds to $\tau_{\text{key}} = 0.75$ and $\tau_{\text{rel}} = 0.75$.

5.2 Main Results

Based on the results in Table 2, a substantial disparity persists between MLLMs and human experts, particularly in process adherence where humans achieve near-perfect consistency (Process F1 0.943) compared to the leading model, Gemini-3-flash-preview (Process F1 0.356). Regarding model performance, closed-source models consistently outperform open-source alternatives, with Gemini-3-flash-preview and GPT-5 demonstrating superior capabilities in both process alignment and numerical stability. However, a critical divergence remains across all evaluated models: they perform significantly better at predicting final scores than replicating granular grading steps. This contrast suggests that current MLLMs largely rely on holistic estimation rather than the precise, step-wise grading of expert assessment.

5.3 OCR Enhancement Impact

To assess the impact of visual clarity on multimodal grading, we conducted a controlled experiment using PaddleOCR-VL(Cui et al., 2025) to extract text and formulas from student responses, providing the OCR output as an auxiliary textual modality

alongside the original image—without altering the visual input—so the model could combine clarified semantics with preserved spatial context. As shown in Table3, this yields only modest gains (e.g., Process F1 for Qwen3-VL-235B-Instruct improves from 0.251 to 0.265), suggesting the main bottleneck lies not in visual recognition noise but in high-level reasoning and evidence-anchored judgment. Such results further indicate that simply reducing text extraction errors is insufficient to achieve human-like grading performance.

5.4 Effectiveness Of RARG

To validate the effectiveness of RARG, we benchmark it against two standard prompting strategies:3-shot , which injects three student responses containing the maximum number of key steps, paired with their corresponding annotations, into the input context as exemplars; and Multi-turn CoT, which decomposes the task into sequential inference turns—identifying steps, deriving rationales, and assigning scores (Brown et al., 2020; Wei et al., 2022; Wang et al., 2023). As presented in Table 3, while all strategies outperform the Zero-shot baseline, RARG achieves the most substantial gains, particularly in grading process metrics. This confirms that retrieving verified pedagogical priors to synthesize adaptive rubrics effectively mitigates parametric hallucinations, ensuring precise, fine-grained assessment.

Quality Assessment of Synthesized Rubrics We prompt GPT-4o (OpenAI et al., 2024) to evaluate synthesized rubrics against expert annotated rubrics across three dimensions (1–5 scale): Accuracy (correctness of constraints), Granularity (detail of step decomposition), and Completeness (coverage of grading points). As illustrated in Table 4, the synthesized rubrics demonstrate superior quality across all metrics. To validate reliability, we randomly sampled 100 rubrics for human expert review. The strong Spearman correlation ($\rho = 0.85$) between human and GPT-4o ratings (averaged across dimensions) confirms the robustness of our evaluation paradigm and substantiates the high pedagogical quality of the generated rubrics.

5.5 Fine-Grained Error Analysis

We propose a fine-grained taxonomy to quantify prediction errors into four distinct categories. Hallucinated Steps refer to fabricating content absent from the context, whereas Misaligned Steps involve generating relevant student response that fails to map to specific ground truth entries. For steps that are correctly matched, errors are classified as either Rationale Mismatch (providing incorrect justification for a correctly located step) and Scoring Deviation (numerical discrepancies that occur only when the underlying rationale is valid).

Model	Acc.	Gran.	Comp.
GPT-5	4.92	4.85	4.78
Gemini-3-Flash-Preview	4.75	4.95	4.68
Qwen3-VL-235B-Instruct	4.80	4.65	4.88

Table 4: Quality assessment of synthesized rubrics. Performance comparison across three dimensions (Acc.: Accuracy, Gran.: Granularity, Comp.: Completeness).

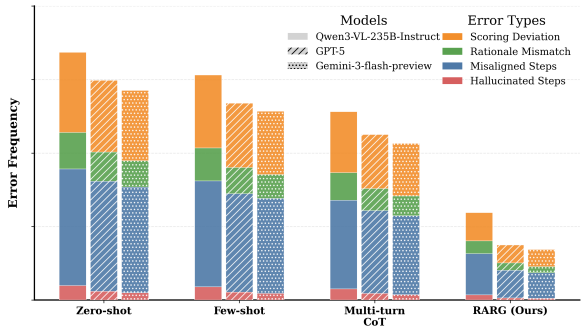

Figure 5: Comparative error analysis across grading strategies.

We apply this taxonomy to analyze the zero-shot performance of Qwen3-VL-235B-Instruct, GPT-5, and Gemini-3-Flash-Preview. To rigorously distinguish ungrounded hallucinations from localization failures, we implement an automated verification pipeline using PaddleOCR-VL and Qwen3-8B-Instruction: text streams are first extracted from raw images via PaddleOCR-VL, after which Qwen3-8B-Instruct verifies whether the predicted steps exist within the extracted context. As shown in Figure 5, the aggregated results reveal that the error distribution is dominated by Misaligned Steps, suggesting that current models lack the granularity required for precise step localization within student responses. Moreover, for successfully aligned steps, the incidence of Scoring Deviation significantly exceeds that of Rationale Mismatch. These patterns collectively highlight a critical disconnect between semantic reasoning and quantitative judgment—while models can often derive valid justifications, they struggle to translate this understanding into accurate numerical scores.

We further investigate several error mitigation strategies to improve grading reliability.As shown

Method	Spear ↑	NMAE ↓	W-F1 ↑	Jacc ↑	Recall ↑	F1 ↑
Base (Zero-shot)	0.419	0.308	0.154	0.104	0.179	0.140
SFT (Final Score, no rubric)	0.545	0.240	0.250	0.225	0.320	0.295
SFT (Final Score, +rubric)	0.615	0.224	0.275	0.285	0.322	0.314
SFT (Process+Score, no rubric)	0.560	0.235	0.260	0.240	0.335	0.310
SFT (Process+Score, +rubric)	0.635	0.215	0.284	0.301	0.396	0.376
RARG (Zero-shot, Ours)	0.650	0.208	0.312	0.334	0.425	0.414
RARG + SFT	0.665	0.198	0.320	0.342	0.432	0.426

Table 5: Performance Comparison between SFT and RARG with Qwen3-VL-4B Backbone

in Figure 5, Multi-turn CoT effectively enhances the logical coherence of reasoning paths but still fails to suppress frequent scoring deviations in the absence of explicit external constraints.In contrast, RARG significantly reduces typical error types, especially misaligned steps and scoring deviations, by anchoring quantitative scoring judgments with standardized, expert-verified educational priors. This demonstrates that constraint-guided reasoning is more critical than mere multi-step generation for accurate assessment.

5.6 Comparison of SFT and RARG

To explore whether fine-tuning on part of the benchmark can improve model grading performance, we conducted Supervised Fine-Tuning (SFT) experiments and compared them with our proposed RARG framework: we collected 11,200 additional mock exam responses with ground-truth scores, used Qwen3-VL-235B to generate step-wise grading rationales as training targets, and fine-tuned Qwen3-VL-4B-Instruct for 1 epoch on 8×A100 GPUs with four supervision signals. Experimental results show that SFT with step-wise process and rubric supervision performs best among pure SFT models (F1=0.376), while our training-free RARG outperforms it (F1=0.414); the combination of RARG and optimal SFT achieves the highest performance (F1=0.426), verifying their synergistic effect. We prioritize RARG for its practicality in K-12 settings, as it adapts to new rubrics dynamically without retraining.

6 Conclusion

We introduce EduMARS, a multimodal K-12 benchmark revealing that current models struggle with precise grading. To address this, we propose RARG, a retrieval-augmented framework that improves scoring reliability by anchoring judgments on expert educational priors.

Limitations

We acknowledge several limitations in this work. First, limited dataset diversity in question types and curricula may restrict generalizability to unseen problems, particularly in highly specialized disciplines with non-standardized formats. Second, heavy dependency on retrieval corpus completeness can lead to performance degradation on rare questions, as the framework requires high-quality educational priors to anchor its reasoning. Furthermore, the extreme variability in authentic student handwriting and complex spatial layouts still pose significant interpretational challenges for multimodal models. Finally, while we focus on scoring precision, broader pedagogical nuances like constructive feedback generation warrant further exploration.

Ethical Considerations

We strictly adhere to ethical guidelines. All data underwent rigorous de-identification to remove PII (e.g., names) to safeguard anonymity. Annotators were professional teachers hired under formal contracts, ensuring fair compensation above local minimum wages and reasonable working hours.

Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive feedback. We also acknowledge the use of generative AI tools in the preparation of this work; specifically, large language models were utilized to assist in professional language polishing and drafting, while AI-based image generation tools provided support for the conceptual design of certain figures. All AI-generated content was rigorously reviewed and

edited by the authors to ensure academic integrity and technical accuracy.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Sami Baral, Lucy Li, Ryan Knight, Alice Ng, Luca Soldainin, Neil Heffernan, and Kyle Lo. 2024. Drawedumath: Evaluating vision language models with expert-annotated students' hand-drawn math images. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Handong Zheng, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. Paddleocr-vl: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model. *Preprint*, arXiv:2510.14528.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. 2025. Omni-MATH: A universal olympiad level mathematic benchmark for large language models. In *The Thirteenth International Conference on Learning Representations*.

Google Gemini Team. 2025. Gemini 3 flash model card. Technical report, Google DeepMind.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. pages 6300–6308.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Peichao Lai, Kexuan Zhang, Yi Lin, Linyihan Zhang, Feiyang Ye, Jinhao Yan, Yanwei Xu, Conghui He, Yilei Wang, Wentao Zhang, and Bin Cui. 2025a. Sasbench: A fine-grained benchmark for evaluating short answer scoring with large language models. *Preprint*, arXiv:2505.07247.

Peichao Lai, Kexuan Zhang, Yi Lin, Linyihan Zhang, Feiyang Ye, Jinhao Yan, Yanwei Xu, Conghui He, Yilei Wang, Wentao Zhang, and 1 others. 2025b. Sasbench: A fine-grained benchmark for evaluating short answer scoring with large language models. *arXiv preprint arXiv:2505.07247*.

Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, Linda Ruth Petzold, Stephen D. Wilson, Woosang Lim, and William Yang Wang. 2025. MMSci: A dataset for graduate-level multi-discipline multimodal scientific understanding.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. 2024a. Criticbench: Benchmarking llms for critique-correct reasoning. *arXiv preprint arXiv:2402.14809*.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. 2024b. CriticBench: Benchmarking LLMs for critique-correct reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1552–1587, Bangkok, Thailand. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

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 *The Twelfth International Conference on Learning Representations*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Gpt-5 system card. Technical report, OpenAI.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Huanxin Sheng, Xinyi Liu, Hangfeng He, Jieyu Zhao, and Jian Kang. 2025. Analyzing uncertainty of LLM-as-a-judge: Interval evaluations with conformal prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11297–11339, Suzhou, China. Association for Computational Linguistics.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating LLM-based judges. In *The Thirteenth International Conference on Learning Representations*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with the math-vision dataset. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024b. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024c. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2025. Self-preference bias in LLM-as-a-judge.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations*.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Processbench: Identifying process errors in mathematical reasoning. In *The 63rd Annual Meeting of the Association for Computational Linguistics*.

Zihao Zhou, Shudong Liu, Maizhen Ning, Wei Liu, Jindong Wang, Derek F Wong, Xiaowei Huang, Qiufeng Wang, and Kaizhu Huang. 2024. Is your model really a good math reasoner? evaluating mathematical reasoning with checklist. *arXiv preprint arXiv:2407.08733*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2024. JudgeLM : Fine-tuned large language models are scalable judges.

Appendix

A.Details of EduMARS

A.1 Data Collection

To ensure the pedagogical rigor and reliability of our dataset, we recruited 10 professional high school teachers, each possessing extensive teaching experience and a minimum of three years in active service. Adhering to ethical labor standards, all experts were compensated at rates consistent with local professional benchmarks. The annotation workflow commenced with the formulation of comprehensive scoring rubrics. Crucially, experts were tasked with anticipating the full spectrum of possible student responses, encompassing not only standard solutions but also alternative valid methods and common misconceptions, rather than relying solely on a fixed reference answer. Subsequently, guided by these robust rubrics and standard reference contexts, the experts conducted a meticulous evaluation of the student submissions to ensure precise and consistent annotation; notably, any response deemed irrelevant to the rubric criteria was explicitly recorded as containing no key steps.

A.2 Annotation Pipeline

To ensure the pedagogical rigor and reliability of our dataset, we recruited 10 professional high school teachers, each possessing extensive teaching experience and a minimum of three years in active service. Adhering to ethical labor standards, all experts were compensated at rates consistent with local professional benchmarks. The annotation workflow commenced with the formulation of comprehensive scoring rubrics. Crucially, experts were tasked with anticipating the full spectrum of possible student responses, encompassing not only standard solutions but also alternative valid methods and common misconceptions, rather than relying solely on a fixed reference answer. Subsequently, guided by these robust rubrics and standard reference contexts, the experts conducted a meticulous evaluation of the student submissions to ensure precise and consistent annotation.

A.3 Quality Control

To ensure consistency, we recruited expert reviewers with teaching expertise comparable to the original annotators and implemented a strict dual-review protocol. Each assessment sample was independently verified by two distinct experts. To minimize bias, a sample was discarded only when both reviewers reached a consensus regarding its invalidity (e.g., severe logical inconsistencies). The specific prompt utilized for the preceding LLM-based rationale standardization is detailed in Table A1. This rigorous verification process ultimately yielded the final high-quality dataset of 4,501 samples.

A.4 Implementation Details of Evaluation

To accurately evaluate the grading process, we align the generated steps $\hat{\mathcal{S}}_{\text{step}}$ with the ground truth $\mathcal{S}_{\text{step}}$ by searching for the optimal correspondence. Regarding key segments, we enforce exact matching for STEM subjects and utilize semantic similarity ($\tau_{\text{key}} = 0.75$) for Liberal Arts. However, for grading rationales across all subjects, we universally employ semantic similarity computed by the Qwen3-embedding-4B model, with a unified threshold of $\tau_{\text{rat}} = 0.75$. This embedding model was selected for its specific adaptability to grading tasks, superior performance, and optimal balance between effectiveness and computational efficiency. Notably, since the steps originate from unique student responses, each predicted step inherently corresponds to at most one ground truth step. We validated the reliability of this alignment by submitting 100 matching results randomly sampled from the output of each tested model to human experts for inspection, which confirmed high consistency with

expert judgment. Finally, regarding score evaluation, since both predicted and ground truth scores are inherently integers, we report the Weighted F1 score to address the natural class imbalance in score distributions. Our experiments were conducted on a server equipped with four NVIDIA A100 GPUs. For the evaluation of ultra-large-scale models, specifically Qwen3-VL-235B-Instruct and GLM-4.6V, we utilized third-party API services provided by the DMXAPI platform to ensure stable inference and consistent throughput.

B. Implementation Details of RARG

B.1 Construction and Retrieval of Specialized Knowledge Base

For the retrieval process, we constructed a professional pedagogical vector database that pairs concise semantic anchors (for efficient indexing) with detailed grading payloads. These payloads are not merely text snippets but encompass a structured repository of granular penalty logic, fine-grained scoring criteria, and concrete historical error examples designed to provide high-fidelity guidance for the generation phase.During the inference stage, the pipeline follows a robust two-stage retrieval and reranking strategy:Initial Retrieval via Semantic Anchors: We utilize the keypoints autonomously generated by the model as independent search queries. To bridge the semantic gap between queries and the database, we employ the Qwen-embedding-4B model to vectorize each keypoint, fetching the Top-N ($N = 5$) candidate rules. These candidates are then aggregated into a high-recall initial pool.Pedagogy-grounded Reranking: To refine the selection, we introduce a Cross-Encoder (BGE-Reranker-Large) to perform a deep semantic interaction between the candidate pool and the problem context. Crucially, this reranking process is strictly conditioned on the relevance between the rules and the [Question, Standard Answer] pair. By intentionally excluding the student's raw response from the reranking query, we effectively decouple the grading standards from the potentially noisy or erroneous student input. This design choice ensures that the final Top-K ($K = 10$) rules injected into the context are fundamentally grounded in the authoritative solution path, thereby filtering out irrelevant pedagogical evidence and maintaining the objectivity of the subsequent step-aware assessment.

Model	Method	Acc. ↑	Gran. ↑	Comp. ↑
GPT-5	Keypoint	**4.92**	**4.85**	**4.78**
	Random	3.12	3.45	2.88
Gemini-3-Flash	Keypoint	**4.75**	**4.95**	**4.68**
	Random	2.95	3.22	2.65
Qwen3-VL-235B	Keypoint	**4.80**	**4.65**	**4.88**
	Random	2.88	3.15	2.72

Table 6: Quality assessment of synthesized rubrics. We compare the rubrics generated via keypoint-based retrieval versus random sampling across three pedagogical dimensions (1-5 scale).

C.Detailed Experiment Results

C.1 Detailed Main Results

As shown in Figure 6, the radar charts reveal three consistent trends across the eight disciplines. First, a clear capability hierarchy exists: the Human Level sets the upper bound, followed by closed-source models, with open-source models exhibiting a stratified performance gap. Second, performance varies significantly by subject; logic-intensive domains like Math and Physics cause notable contraction in agent scores. In contrast, Biology and Chemistry show a balanced, evenly distributed gradient, where model performance closely approaches the Human Level due to the lower complexity of fill-in-the-blank tasks in these disciplines. Finally, a persistent process-outcome gap indicates that models consistently achieve higher scores on Final Score metrics than on Grading Process metrics, reflecting the ongoing challenge of precise reasoning retrieval.

C.2 Ablation Study of the RARG Framework

To disentangle the contributions of the retrieval component and the intermediate rubric generation mechanism within the RARG framework, we conducted a series of controlled experiments. These analyses aim to address three fundamental research questions: (1) Does the two-stage point-wise retrieval strategy effectively provide accurate pedagogical evidence? (2) Does the explicit synthesis of a structured rubric significantly enhance grading precision compared to using raw evidence directly?

Validation of Retrieval Effectiveness via Random Search. To verify that RARG's performance stems from precise knowledge alignment rather than stochastic factors, we conducted an ablation study comparing our keypoint-based retrieval

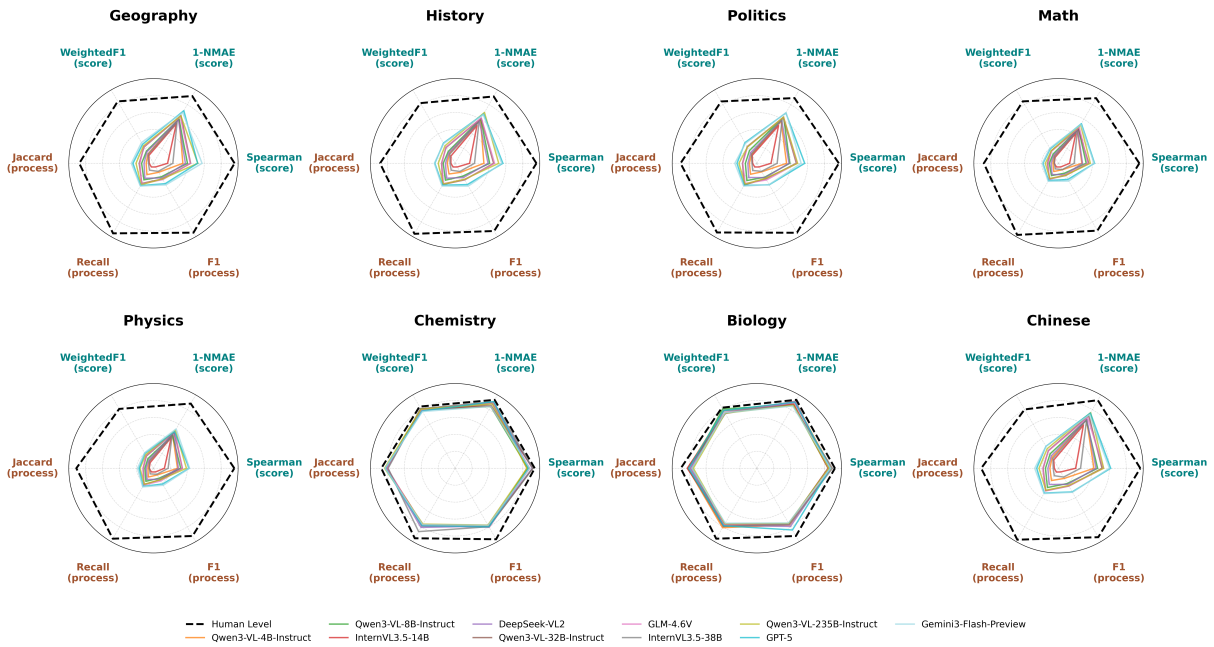

Figure 6: Subject-wise performance breakdown of various models across eight disciplines, illustrating the comparative analysis between AI agents and the human benchmark.

against a Random Search baseline. As illustrated in Table 6, the quality of rubrics synthesized under the Random Search configuration suffers a catastrophic decline across all three pedagogical dimensions. Specifically, for the Qwen3-VL-235B model, Accuracy (Acc.) drops significantly from 4.80 to 2.88, while Completeness (Comp.) plummets to 2.72, indicating that without precise retrieval, the model fails to incorporate essential grading criteria and specific error patterns. Even Granularity (Gran.) sees a sharp reduction to 3.15, as the model reverts to generating generic, non-case-specific feedback lacking the necessary step-wise depth. This sharp contrast confirms that the precision of our two-stage retrieval mechanism is the fundamental pillar for generating high-fidelity, expert-aligned rubrics, rather than mere context enrichment.

Method	Final Score			Grading Process		
	Spr.↑	NMAE↓	WF1↑	Jac.↑	Rec.↑	F1↑
RARG (Ours)	**0.755**	**0.112**	**0.452**	**0.435**	**0.582**	**0.472**
DRA (Raw)	0.635	0.198	0.310	0.292	0.380	0.301

Table 7: Comparison of RARG and DRA on final score and grading process evaluation metrics.

Necessity of Structured Rubric Generation. We further investigated whether the intermediate step of synthesizing raw snippets into a formal rubric is essential for accurate assessment. We compared RARG against a Direct Retrieval-Augmented (DRA) approach, where the model performs grading directly based on the top-10 raw retrieved snippets without the synthesis step. The results reveal that while DRA improves upon the Zero-shot baseline, it consistently lags behind RARG, particularly in the Scoring Deviation metric. This suggests that the generated rubric acts as a critical "logical anchor," reducing the cognitive load on the MLLM by transforming fragmented information into a structured, executable scoring standard.

C.3 Case Study

Figure 7 illustrates a complete data example from the benchmark. Within the 'Answer Image', the key derivation steps are specifically highlighted with orange bounding boxes. These boxes represent the critical milestones identified in the annotations, allowing the system to evaluate the correctness of the student's problem-solving logic and assign scores to individual steps.

Figure 8,Figure9, Figure10,Figure11 present a qualitative evaluation of four distinct grading strategies: Zero-shot, Few-shot (In-context Learning), Multi-turn Chain-of-Thought (MT-CoT), and Retrieval-Augmented Adaptive-Rubric Grading (RARG), where green bounding boxes denote correct logical alignments and red boxes indicate reasoning failures.

The Zero-shot baseline exhibits fundamental

recognition of initial kinematic modeling but suffers from significant logical hallucinations, such as misidentifying the periodic formula as evidence for particle entry direction while failing to verify numerical consistency during velocity synthesis. Although the introduction of Few-shot exemplars enhances structural segmentation, The model still exhibits short-sighted reasoning in multi-step problems, failing to connect correctly identified intermediate formulas to the final displacement requirements. Although MT-CoT greatly improves rationale granularity and aligns well with early solution milestones, it struggles with the semantic gap between symbolic period formulas and specific spatial constraints.In contrast, the RARG framework achieves the most robust performance, as it can precisely locate key derivation steps and assign the corresponding scores correctly. These results confirm that RARG provides the most consistent alignment through explicit reasoning and external knowledge integration.

C.4 case study for rubric synthesis

```
Problem
14. (14 points) As shown in the figure,
    small objects 1 and 2 on a horizontal
    turntable are connected by a light,
    thin string. Their masses are m and 2m
    , respectively. The string is kept
    straight and just barely without
    tension while the objects are at rest
    on the turntable. The turntable can
    rotate around a vertical central axis
    OO'. The coefficient of dynamic
    friction between both objects and the
    turntable surface is μ, and the maximum
     static friction is assumed to be equal
     to the sliding friction. Both objects
    are collinear with the axis O. The
    distance from object 1 to the axis is r,
     and the distance from object 2 to the
    axis is 2r. The acceleration due to
    gravity is g. When the turntable starts
     rotating from rest and the angular
    velocity increases very slowly, solve
    the following questions for this
    process:

1. Find the angular velocity of the
    turntable when the string just begins
    to have tension.
2. Find the friction force acting on object
    1 when the angular velocity of the
    turntable is ω = sqrt(2μg/3r).
3. Find the angular velocity when both
    object 1 and object 2 are thrown off
    the turntable.
```

```
Standard Answer:

(1) When the light string just begins to
    have tension, the friction between
    object 2 and the turntable reaches the
    maximum static friction. According to
    Newton's Second Law:
\mu \cdot 2mg = 2m\omega_0^2 \cdot 2r
Solving for ω0:
\omega_0 = \sqrt{\frac{\mu g}{2r}}

(2) When the angular velocity of the
    turntable is ω = sqrt(2μg/3r) > sqrt(μg/2r), the
    friction between object 2 and the
    turntable is the maximum static
    friction. Then:
For object 2: F_T + μ·2mg = 2mω^2·2r
For object 1: F_T + F_f = mω^2·r
Solving these equations gives:
F_f = 0

(3) When both objects are about to be thrown
     off, the friction on object 1 reaches
    its maximum value in the direction away
     from the center (to oppose the tension
    ).
For object 2: F'_T + μ·2mg = 2mω_1^2·2r
For object 1: F'_T − μmg = mω_1^2·r
Solving for ω1:
\omega_1 = \sqrt{\frac{\mu g}{r}}
```

Table8 and Table9 show the scoring rubrics generated by human annotators and by Gemini 3 Flash Preview, respectively, for the problem and its standard answer presented in the code block, illustrating the quality of the model-generated rubric.

D. Prompts

Annotation Process: Prompt for Eliminating Bias and Extracting Objective Evaluation

```
Role: Senior K-12 Education Assessment
    Expert.
Task: Compare the question, standard
    answer, and the teacher's original
    rubric to generate a standardized
    marking rubric for students' answers.

Requirements:
1. De-biasing: Eliminate all subjective
    emotional words (e.g.,
    "unfortunately", "careless") and
    retain only fact-based judgments.
2. Standardization: Convert colloquial
    evaluations into objective academic
    terms.

Input Format:
Question: {Question}
Standard_Answer: {Standard Answer}
Teacher_Raw_Comment: {Original Comment}
```

Evaluation Process: Zero-shot

```
Task: Act as an expert grader and assign
     scores to the student's handwritten
     response by analyzing the image in the
     context of the question and standard
     answer.

Requirements:
1. Score directly from the student's answer
     image, do not assume or hallucinate
     content.
2. Identify all essential reasoning or
     computation steps implied by the
     question and standard answer.
3. For each step, provide a factual
     rationale that explicitly references
     evidence in the student's response and
     aligns with the expected solution.

Input:
Question: {Question}
Standard Answer: {Standard Answer}
Student_Image: [Student Answer Sheet Image]

Output:
(Key Step: {Description} | Rationale:
     {Evidence-based justification} |
     Score: {Numeric Score}) | ... | Total
     Score: {Total}
```

Evaluation Process: MT-CoT Step 1 (Transcription & Key Steps Generation)

```
Task: Extract and structure the student's
     handwritten solution into key
     problem-solving steps by aligning it
     with the question and the standard
     answer.

Requirements:
1. Maintain consistent formula formatting.
2. Clearly separate given problem context
     from student-generated content.
3. Identify and sequence only those steps
     that correspond to essential reasoning
     or computation stages in the standard
     solution.

Input:
[Student Answer Sheet Image], Question:
     {Question},
Standard Answer: {Standard Answer}

Output: Key Step 1: {Content} | Key Step 2:
     {Content} | ...
```

Evaluation Process: MT-CoT Step 2 (Rationale Generation)

```
Task: Generate scoring rationales for the
     student's key steps by aligning them
     with the question and the standard
     answer and strudent response.

Requirements:

1.Analyze each student key step in the
     context of the original response.
Determine whether the step correctly
     addresses the required reasoning or
     computation.
2.Produce a clear [Key Rationale] for each
     step, explicitly justifying its
     correctness, partial credit, or error
     based on alignment with the expected
     solution.
Input:
[Student Answer Sheet Image],
Question: {Question};
Standard Answer: {Standard Answer};
     Student_Key_Steps: {Results of Step 1}

Output: Key Step {ID}: [Key Rationale]|...
```

Evaluation Process: MT-CoT Step 3 (Final Score Aggregation)

```
Task: Assign a final score by synthesizing
     the key steps, their corresponding
     rationales, and alignment with the
     question and standard answer.

Requirements:
1. Review each Key Step and its [Key
     Rationale] to determine an appropriate
     partial score.
2. Ensure scoring is grounded in the
     original student response, not
     inferred content.
3. Sum individual step scores to produce a
     coherent total score consistent with
     expert grading practice.

Input:
Question: {Question};
Standard Answer: {Standard Answer};
[Student Answer Sheet Image];
     Student_Key_Steps: {Results of Step
     1}; Key_Rationales: {Results of Step 2}

Output: Key Step 1: {Score} Key Step 2:
     {Score} ... Total Score: {Total}
```

Evaluation Process: Few-shot Version

Assessment Process: 3-Shot In-Context Grading

```
Task: Act as an expert grader and assign
     scores to the student's handwritten
     response by analyzing the image in the
```

```
    context of the question and standard
    answer.

Requirements:
1. Score directly from the student's
    answer-do not assume or hallucinate
    content.
2. Identify all essential reasoning or
    computation steps implied by the
    question and standard answer.
3. For each step, provide a factual
    rationale that explicitly references
    evidence in the student's response.

Input Format:
Question: {Question}
Standard Answer: {Standard Answer}
Student_Image: [Transcribed content from
    the student's handwritten response]
[EXAMPLES:{example}]
Output Format:
(Key Step: {Description} | Rationale:
    {Evidence-based justification} |
    Score: {Numeric Score}) | ... | Total
    Score: {Total}
```

RARG Process:Keypoint Extraction

```
Task: Extract the key semantic points
    (Keypoints) that must be verified
    during scoring from the question and
    standard answer.

Requirements:
1. Identify essential logical or
    computational elements required to
    solve the problem.
2. Include formula preconditions,
    intermediate state variables, and
    final results.
3. Express each Keypoint as a concise,
    self-contained factual statement.

Input:
Question: {Question}
Standard Answer: {Standard Answer}

Output:
List of Keypoints: 1. {Keypoint}; 2.
    {Keypoint}; ...
```

RARG Process: Adaptive Rubric Synthesis

```
Task: Synthesize an adaptive scoring guide
    from the extracted Keypoints and
    grading prior knowledge.

Requirements:
1. For each Keypoint, define a clear
    scoring criterion that specifies what
    constitutes correct, partially
    correct, or incorrect responses.
2. Incorporate fault tolerance by
    recognizing mathematically or
    semantically equivalent student
    expressions.
3. Use the grading prior (e.g., common
    error patterns, valid alternative
    formulations) to refine judgment
    boundaries.

Input:
Keypoints: {List of essential semantic
    points from the question and standard
    answer}
Grading_Prior: {Explanations of common
    errors, acceptable variants, or
    domain-specific equivalences}

Output:
[No.]: {Step ID} | [Steps]: {Keypoint
    Summary} | [Criteria]: {Refined,
    evidence-aware scoring condition} |
    [Max Score]: {Integer}
```

RARG Process: Rubric-Guided Grading

```
Task: Grade the student's handwritten
    response using the enhanced scoring
    guide, simulating expert reasoning.

Requirements:
1. Analyze the student's answer image in
    the context of the question and
    standard answer.
2. For each Keypoint in the Enhanced
    Rubric, verify whether the student's
    response satisfies the refined
    criteria.
3. Assign a score per Keypoint with a
    rationale that cites specific evidence
    from the student's work and references
    the corresponding criterion.

Input:
Question: {Question}
Standard Answer: {Standard Answer}
Student_Image: [Student Answer Sheet Image]
Keypoints: {List of essential semantic
    points}
Enhanced_Rubric: [No.]: {Step ID} |
    [Steps]: {Keypoint Summary} |
    [Criteria]: {Refined scoring
    condition} | [Max Score]: {Integer}

Output:
(Key Step: {Keypoint Summary} | Rationale:
    {Evidence-based justification aligned
    with the Enhanced Rubric} | Score:
    {Numeric Score}) | ... | Total Score:
    {Total}
```

Verify Hallucination

```
Role: Marking Auditor
```

```
Task: Determine whether the AI-generated
     scoring report contains hallucinations
     by verifying its claims against the
     student's actual key steps (text
     transcription).

Requirements:
1. Every factual claim in the Rationale
     (e.g., presence/absence of a value,
     formula, or reasoning step) must be
     supported by the provided student key
     steps.
2. No inference beyond the given text is
     allowed.

Input:
Student_Key_Steps: {Text transcription of
     the student's solution steps}
OCR_TEXT: {OCR_TEXT}

Output:
Hallucination_Detected: [Yes / No]
```

Quality Assessment of Synthesized Rubrics

```
Task: Evaluate the quality of a synthesized
     scoring rubric by comparing it to an
     expert-annotated reference rubric.

Instructions:
Rate the synthesized rubric on the
     following three dimensions using a
     1---5 integer scale (1 = very poor, 5
     = excellent). Base your judgment
     solely on the content of the two
     rubrics.

Dimensions:
1. Accuracy: Does the synthesized rubric
     correctly capture the logical
     constraints, conditions, and scoring
     logic of the expert rubric? Are there
     factual errors or misinterpretations?
2. Granularity: Does the synthesized rubric
     decompose the solution process into
     steps with a level of detail
     comparable to the expert rubric? Are
     intermediate reasoning stages
     adequately specified?
3. Completeness: Does the synthesized
     rubric cover all essential grading
     points present in the expert rubric?
     Are any critical criteria missing?

Input:
- Expert_Rubric: {Step-by-step
     expert-annotated scoring guide}
- Synthesized_Rubric: {Automatically
     generated scoring guide}

Output Format (strictly follow):
Accuracy: [1-5] | Granularity: [1-5] |
     Completeness: [1-5]
```

No.	Steps	Scoring Point Description	Score	Cumulative Score
Critical State: Initiation of Tension (3 pts)				
1	Critical Analysis	Identify that as ω increases, Object 2 reaches its maximum static friction first due to its larger radius and mass.	1 pt	1 pt
2	Centripetal Equation	Set up the equation for Object 2: $\mu(2m)g = 2m\omega_0^2(2r)$.	1 pt	2 pts
3	Calculation	Solve for $\omega_0 = \sqrt{\frac{\mu g}{2r}}$.	1 pt	3 pts
Equilibrium Analysis of Intermediate State (5 pts)				
4	State Determination	Compare $\omega = \sqrt{\frac{2\mu g}{3r}}$ with ω_0. Since $\omega > \omega_0$, determine that the string has tension and Object 2 is at max friction.	1 pt	4 pt
5	Dual-Object Modeling	For Object 2: $F_T + \mu(2m)g = 2m\omega^2(2r)$. For Object 1: $F_T + F_f = m\omega^2 r$.	3 pts	7 pt
6	Result	Solve the system to find friction on Object 1: $F_f = 0$.	1 pt	8 pt
Limit State: Both Objects Thrown Off (6 pts)				
7	Friction Direction	Key Point: Identify that when about to be thrown off, Object 1's static friction is at its maximum and points away from the center.	2 pts	10 pt
8	Limit Equations	Establish the system: $F_T' + 2\mu mg = 2m\omega_1^2(2r)$ and $F_T' - \mu mg = m\omega_1^2 r$.	2 pts	12 pt
9	Final Solution	Eliminate F_T' to solve for the final critical angular velocity: $\omega_1 = \sqrt{\frac{\mu g}{r}}$.	2 pts	14 pt

Table 8: Teacher's Rubric

No.	Steps	Scoring Point Description	Score	Cumulative Score
Angular Velocity when Tension Begins (3 points)				
1	Critical State Analysis	Identify that when tension just begins, the static friction of Object 2 reaches its maximum value: $f_{max} = \mu(2m)g$.	2pts	2 pts
2	Solve Equation	Set up Newton's Second Law: $\mu 2mg = 2m\omega_0^2 \cdot 2r$ and solve for $\omega_0 = \sqrt{\frac{\mu g}{2r}}$.	1 pt	3 pts
Friction at Specific Angular Velocity (5 points)				
3	State Determination	Compare $\omega = \sqrt{\frac{2\mu g}{3r}}$ with ω_0. Since $\omega > \omega_0$, Object 2 is at max friction and the string has tension.	1 pt	4 pts
4	Force Eq: Object 2	Establish equation for Object 2: $F_T + \mu 2mg = 2m\omega^2 \cdot 2r$.	2 pts	6 pts
5	Force Eq: Object 1	Establish equation for Object 1: $F_T + F_f = m\omega^2 \cdot r$.	1 pt	7 pts
6	Final Calculation	Solve the system to find friction on Object 1: $F_f = 0$.	1 pt	8 pts
Angular Velocity when Thrown Off (6 points)				
7	Critical State Analysis	Note that when both are about to be thrown off, Object 1 also reaches max friction, pointing away from the center.	2 pts	10 pts
8	System of Equations	Set up equations: $F_T' + \mu 2mg = 2m\omega_1^2 \cdot 2r$ and $F_T' - \mu mg = m\omega_1^2 \cdot r$.	2 pts	12 pts
9	Solve for ω_1	Eliminate F_T' to solve for the final angular velocity: $\omega_1 = \sqrt{\frac{\mu g}{r}}$.	2 pts	14 pts

Table 9: Gemini-3-Flash-Preview Generated Rubric

Problem:As shown in the figure, a uniform electric field with an intensity of $E = 800\ \text{N/C}$ directed along the negative y-axis exists above the x-axis. A uniform magnetic field with a magnetic flux density of $B = 0.04\ \text{T}$ directed perpendicularly into the paper exists below the x-axis. A positively charged particle with a specific charge of $\frac{q}{m} = 1.0 \times 10^{7}\ \text{C/kg}$ is injected from position $P(0, 0.1\ \text{m})$ on the y-axis with an initial velocity $v_{0} = 4 \times 10^{4}\ \text{m/s}$ along the positive x-axis into the uniform electric field. Neglecting the gravity of the particle,
find:(1) The magnitude and direction of the velocity v when the particle first enters the magnetic field;
(2) The total time elapsed from the moment the particle is emitted from point P until it crosses the x-axis for the second time, and the distance from the origin O to the position where it crosses the x-axis for the second time.

Standard Answer:
(1) The particle moves from point P until it crosses the x-axis for the first time, exhibiting projectile-like motion. Upon reaching the x-axis, let the velocity component in the direction of the electric field be vy, and the angle between the resultant velocity v and the positive x-axis be theta. We have: qE = ma (1 pt) vy^2 = 2ay (1 pt) v = sqrt(v0^2 + vy^2) (1 pt) tan theta = vy / v0 (1 pt) Solving the above equations simultaneously: v = 4*sqrt(2) * 10^4 m/s, theta = 45 degrees (2 pts)

(2) Let t1 be the time taken to travel from P to the x-axis for the first time, then: y = (vy / 2) * t1, solving gives t1 = 5 * 10^-6 s (1 pt) Moving at a constant speed in the v0 direction, its displacement is x1 = v0 * t1 = 0.2 m (1 pt) Entering the magnetic field and performing uniform circular motion, we have qvB = m * (v^2 / r), solving gives r = sqrt(2) / 10 m When crossing the x-axis for the second time, the corresponding central angle of its trajectory is 90 degrees, and the chord length is L = sqrt(2) * r = 0.2 m (1 pt) The motion time is t2 = (2 * pi * r) / (4 * v), solving gives: t2 = 1.25 * pi * 10^-6 s (1 pt) In summary, we obtain t = t1 + t2 = (5 + 1.25 * pi) * 10^-6 s, d = x1 + L = 0.4 m (1 pt)

Full Score:11


```
{
  "Annotation": {
      "step": "H = 1/2 * g * t^2, vy = gt = sqrt(2gH)",
      "rationale": "The student correctly modeled the projectile-like motion in the electric field, using displacement and time relations to derive the vertical velocity component.",
      "score": 2
    },
    {
      "step": "V = sqrt(V0^2 + Vy^2) = 4*sqrt(2) * 10^4 m/s",
      "rationale": "The student accurately applied the velocity synthesis principle and performed the calculation correctly to find the resultant velocity magnitude.",
      "score": 2
    },
    {
      "step": "Angle with x-axis is 45 degrees",
      "rationale": "The student correctly determined the direction of the velocity vector at the point of entry into the magnetic field.",
      "score": 2
    },
  "overall_score": 6
}
```


Answer Image

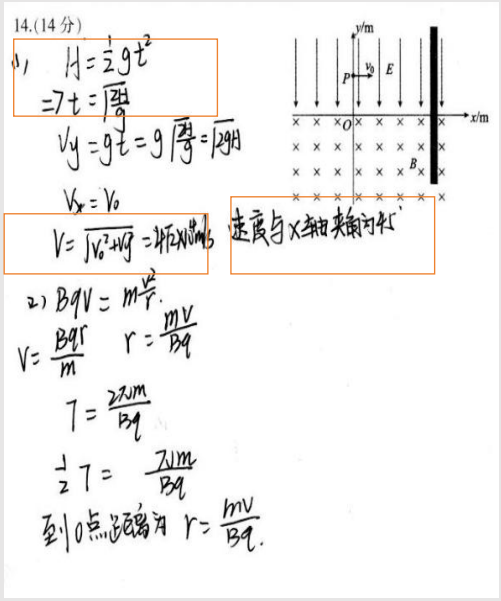

Figure 7: An illustrative example of EduMARS

```
{
  "Model_Generate_Zero_Shot": {
      "step": "H = 1/2 * g * t^2, vy = gt = sqrt(2gH)",
      "rationale": "Correct derivation of $v_y$ via electric field kinematic modeling.",
      "score": 2
    },
    {
      "step": "V = sqrt(V0^2 + Vy^2) = 4*sqrt(2) * 10^4 m/s,Angle with x-axis is 45 degrees",
      "rationale": "Valid application of the Pythagorean theorem for velocity synthesis; resultant magnitude calculation is verified as correct based on previous motion decomposition.",
      "score": 2
    },
    {
      "step": "$\frac{1}{2}T = \frac{\pi m}{Bq}$",
      "rationale": "The student correctly determined the direction of the velocity vector at the point of entry into the magnetic field.",
      "score": 1
    },
   "overall_score": 5
}
```


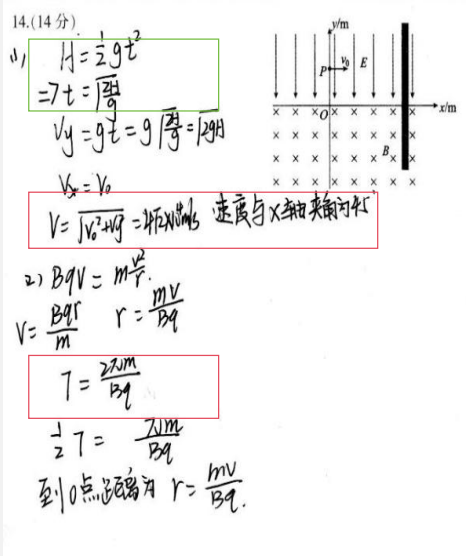

Figure 8: An illustrative example of Gemini-3-Flash-Preview output under the zero-shot setting, corresponding to the case shown in 7.


```
{
  "Model_Generate_3_Shot": [
    {
      "step": "H = 1/2 * g * t^2, vy = gt = sqrt(2gH)",
      "rationale": "The student correctly established the kinematic model for projectile-like motion in the electric field to derive the vertical velocity component.",
      "score": 2
    },
    {
      "step": " Angle with x-axis is 45 degrees",
      "rationale": "Accurate application of the velocity synthesis principle and correct determination of the resultant velocity's direction.",
      "score": 4
    },
    {
      "step": "Bqv = m * (v^2 / r)",
      "rationale": "Proper identification of the Lorentz force acting as the centripetal force for uniform circular motion in the magnetic field.",
      "score": 1
    },
    {
      "step": "T = 2 * pi * m / Bq",
      "rationale": "Correct use of the cyclotron period formula to describe the particle's periodic motion properties within the magnetic field.",
      "score": 1
    }
  ],
  "overall_score": 8
}
```


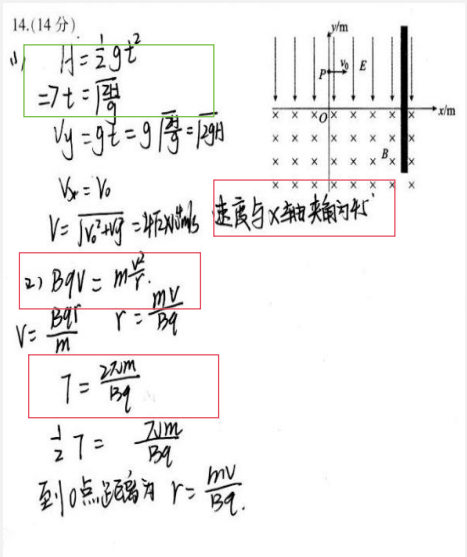

Figure 9: An illustrative example of Gemini-3-Flash-Preview output under the 3-shots setting, corresponding to the case shown in 7.

```
{
 "Model_Generate_MT_Cot": [
   {
     "step": "H = 1/2 * g * t^2, vy = gt = sqrt(2gH)",
     "rationale": "Correctly modeled the vertical motion in the electric field to derive the vertical velocity component via kinematic relations.",
     "score": 2
   },
   {
     "step": "V = sqrt(V0^2 + Vy^2) = 4*sqrt(2) * 10^4 m/s",
     "rationale": "Accurately applied the principle of velocity vector synthesis and performed the calculation correctly to find the resultant magnitude.",
     "score": 2
   },
   {
     "step": "Angle with x-axis is 45 degrees",
     "rationale": "The student successfully determined the direction of the velocity vector at the point of entry into the magnetic field.",
     "score": 2
   },
   {
     "step": "T = 2*pi*m / Bq",
     "rationale": "Properly identified the fundamental formula for the cyclotron period, demonstrating an understanding of the particle's motion in the magnetic field.",
     "score": 1
   }
 ],
 "overall_score": 7
}
```


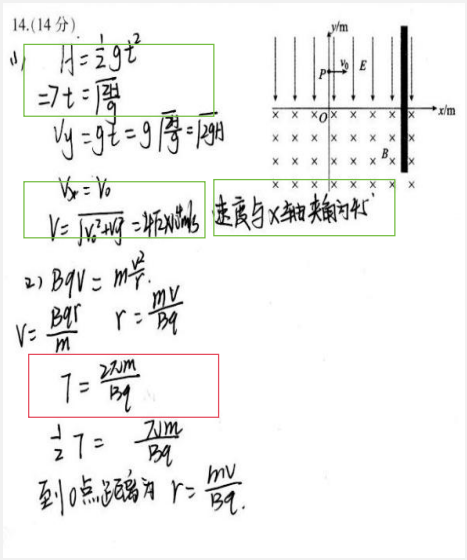

Figure 10: An illustrative example of Gemini-3-Flash-Preview output under the multi-turn cot setting, corresponding to the case shown in 7.


```
{
 "Model_Generate_RARG": [
   {
     "step": "H = 1/2 * g * t^2, vy = gt = sqrt(2gH)",
     "rationale": "The student successfully calculated the vertical velocity by analyzing the motion within the electric field.",
     "score": 2
   },
   {
     "step": "V = sqrt(V0^2 + Vy^2) = 4*sqrt(2) * 10^4 m/s",
     "rationale": "The total velocity magnitude was correctly found using the vector synthesis method.",
     "score": 2
   },
   {
     "step": "Angle with x-axis is 45 degrees",
     "rationale": "The deflection angle was accurately determined based on the velocity components.",
     "score": 2
   }
 ],
 "overall_score": 6
}
```


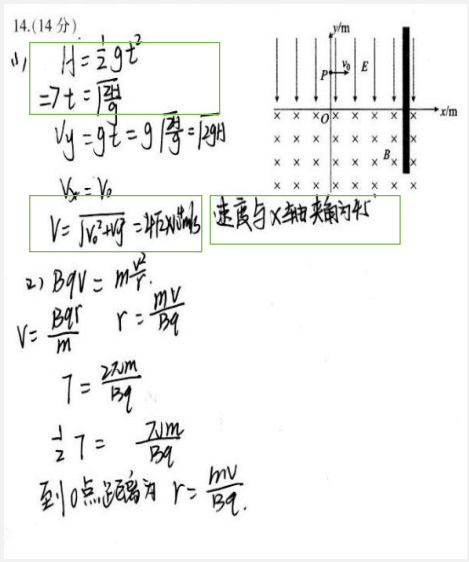

Figure 11: An illustrative example of Gemini-3-Flash-Preview output under the RARG, corresponding to the case shown in 7.

Figure 12: Annotation page for rubric

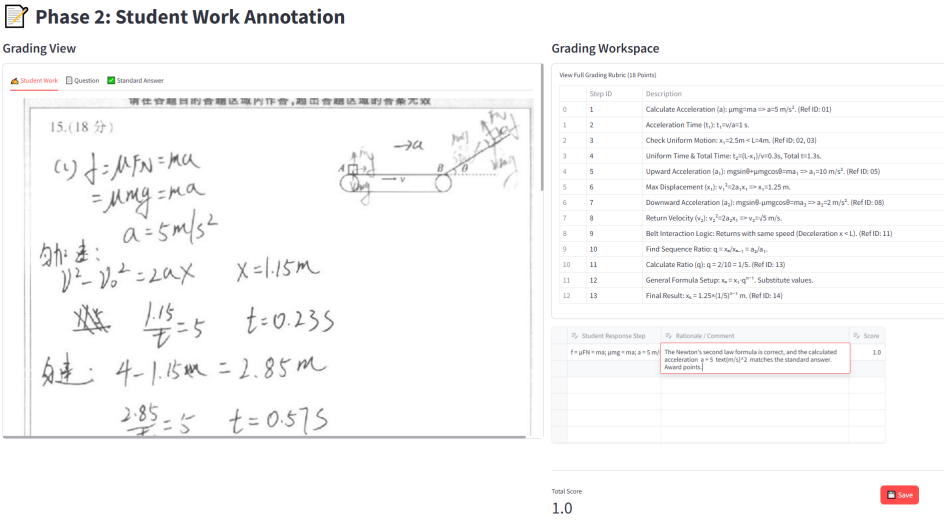

Figure 13: Annotation page for response image