Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Special Track on Human-Centred AI

Enhancing Automated Grading in Science Education through LLM-Driven Causal Reasoning and Multimodal Analysis

Haohao Zhu[1] , **Tingting Li**[2] , **Peng He**[2] , **Jiayu Zhou**[3]
[1]Michigan State University
[2]Washington State University
[3]University of Michigan
zhuhaoha@msu.edu, {tingting.li1, peng.he}@wsu.edu, jiayuz@umich.edu

Abstract

Automated assessment of open responses in K–12 science education poses significant challenges due to the multimodal nature of student work, which often integrates textual explanations, drawings, and handwritten elements. Traditional evaluation methods that focus solely on textual analysis fail to capture the full breadth of student reasoning and are susceptible to biases such as handwriting neatness or answer length. In this paper, we propose a novel LLM-augmented multimodal evaluation framework that addresses these limitations through a comprehensive, bias-corrected grading system. Our approach leverages LLMs to generate causal knowledge graphs that encapsulate the essential conceptual relationships in student responses, comparing these graphs with those derived automatically from the rubrics and submissions. Experimental results demonstrate that our framework improves grading accuracy and consistency over deep supervised learning and few-shot LLM baselines.

1 Introduction

Evaluating open-ended responses in K-12 education remains a complex task due to the varied ways students articulate their understanding [Li *et al.*, 2023b; Li *et al.*, 2023d]. This is a particular challenge in science education, where the new vision of science education is to develop usable knowledge for students [States, 2013]. Students with usable knowledge should be able to apply their knowledge to solve complex problems or explain real-world phenomena [Li *et al.*, 2023a; He *et al.*, 2023]. Scientific modeling has been recognized as a high-leverage practice to engage students in science learning [Li, 2024; Li *et al.*, 2023c]. Rooted in the theory of multimodality [Kress, 2009; Ouyang *et al.*, 2022], scientific modeling can engage students in a richer and more effective learning experience, in which they can process information[Wang *et al.*, 2019] more effectively through multiple sensory channels such as visual, auditory, and tactile [Kress, 2009]. Consequently, scientific models are multi-representations (e.g., graphs, text-based explanations) constructed to explain or make predictions about natural phenomena by identifying critical components and relations between those components using evidence [Schwarz *et al.*, 2017]. Multi-representations offer students with sensory limitations (e.g., auditory or visual) opportunities to learn and express ideas through other channels.

Despite the importance of modeling, the development of scientific models challenges young students and teachers, especially at the elementary level [Li *et al.*, 2023c]. This is in part due to the complexity and diversity that constructed models bring to model evaluation [Li *et al.*, 2024; Li *et al.*, 2023b]. Elementary students often struggle to articulate their understanding clearly, resulting in answers that can combine text with drawings. These multimodal elements frequently contain critical conceptual information, even when responses are misspelled, syntactically unconventional, or partially illegible [Li *et al.*, 2024; Li *et al.*, 2023d]. Traditional single-modal models, which predominantly rely on textual input, struggle to process such heterogeneous inputs [Li *et al.*, 2023b; Ouyang *et al.*, 2023]. While multimodal models provide a broader analytical scope, they often face limitations in distinguishing between unconventional yet valid reasoning and superficially well-formed but conceptually flawed answers [Chia *et al.*, 2024; Wang *et al.*, 2020].

Recent advances in Large Language Models (LLMs) have introduced transformative avenues for the interpretation and evaluation of both textual and visual inputs [Lu *et al.*, 2023]. For example, models such as GPT-4 have been effectively employed to grade handwritten university-level mathematics responses with commendable initial accuracy [Caraeni *et al.*, 2025], and—when appropriately prompted—they are capable of furnishing both scores and detailed explanations for middle-school science answers [Cohn *et al.*, 2024]. These developments suggest the potential for LLMs to mediate the nuanced challenges inherent in K-12 assessment, wherein student responses often manifest as multifaceted representations that extend beyond conventional textual form. However, the direct application of LLMs to the evaluation of scientific models has two outstanding issues: first, human evaluators may inadvertently introduce biases influenced by factors such as handwriting neatness, answer length, or preconceived notions—factors that compromise grading fairness and consistency [Cohn *et al.*, 2024; Botelho *et al.*, 2023]; and second, the comprehensive, multimodal nature of many grading

*Code link: https://github.com/illidanlab/llm_grading_edu.git.

rubrics challenges zero-shot or few-shot LLM paradigms to capture the full spectrum of student reasoning.

To address these challenges, we propose an LLM-augmented multimodal evaluation framework that first calibrates grading biases before leveraging LLM capabilities to construct a causal reasoning graph for assessment. In our approach, an LLM is employed to analyze textual responses and generate an expected *reasoning graph* that encapsulates key conceptual elements; this graph is subsequently compared with one automatically derived from the student's answer. Our framework evaluates the clarity of handwriting and the informativeness of any accompanying drawings or diagrams. By integrating visual analysis, handwriting recognition, and representations from reasoning graph, our method provides a holistic assessment mechanism that captures the diverse modalities through which students articulate their understanding. Our approach is designed to mitigate superficial biases and to ensure that evaluations are anchored in the conceptual robustness of student responses.

Contributions. Our contributions are as follows:

- We present a novel multimodal framework for automated assessment in elementary science education that systematically fuses LLM-based text analysis, visual drawing evaluation, and knowledge graph reasoning to address the challenges posed by multi-representational student work.
- We introduce a bias correction mechanism that leverages handwriting clarity and drawing quality metrics as auxiliary inputs, thereby calibrating human scores and mitigating the impact of subjective presentation biases.
- Through comprehensive empirical studies, we show that our multimodal approach enhances grading accuracy and consistency relative to deep supervised learning and few-shot LLM baselines, with each modality contributing to the overall robustness of the grading.

2 Related Work

Automated Grading and LLMs in Education. Automated short-answer grading (ASAG) has been studied for decades. Early systems like *e-rater* focused on essay scoring using hand-crafted features and rubrics [Burstein *et al.*, 2013]. Traditional ASAG methods often used lexical overlap or keyword matching against model answers, which limited their ability to assess deeper understanding [Sultan *et al.*, 2016]. More recent approaches employed machine learning and deep neural networks to predict scores from text, achieving improved accuracy by learning from large corpora of student responses [Bonthu *et al.*, 2021]. Nonetheless, these models typically treat the grading task as a pure text regression or classification problem, lacking the ability to explain their decisions or account for non-textual cues.

The advent of large-scale pre-trained language models has opened new horizons for ASAG. Transformer-based models fine-tuned on grading tasks have outperformed earlier methods in benchmarks [Sung *et al.*, 2019]. Furthermore, few-shot and zero-shot prompting with LLMs (like GPT-3.5 or GPT-4) have shown that these models can approximate scoring without task-specific training [Jiang and Bosch, 2024]. For instance, [Cohn *et al.*, 2024] used GPT-4 with chain-of-thought prompts to evaluate middle-school science answers, enabling the model to provide a rationale alongside a score.

Such capabilities are valuable in educational settings where explaining the grade is almost as important as the grade itself. Our work builds on this line of research by using an LLM to not only score but also generate structured knowledge graphs representation for deeper comparison and enabling human-in-the-loop capability.

LLM-Based Knowledge Graph. Recent advancements in Large Language Models (LLMs) have significantly improved the construction and application of knowledge graphs in educational assessments [Abu-Rasheed *et al.*, 2025]. LLMs can extract entities and relationships from text, enabling automated generation of structured representations of educational content [Bui *et al.*, 2024].

Beyond entity extraction, LLMs also facilitate causal reasoning, enhancing knowledge representation. [Abdulaal *et al.*, 2024] introduced *Causal Modelling Agents (CMA)*, integrating LLM-driven reasoning with Deep Structural Causal Models (DSCMs) for improved causal discovery. Similarly, [Khatibi *et al.*, 2024] proposed *Autonomous LLM-Augmented Causal Discovery (ALCM)*, combining data-driven and LLM-based causal inference. Building on these advancements, we generate *causal knowledge graphs* for both expected and student answers, enabling a *semantic alignment-based* grading methodology that assesses conceptual correctness beyond text similarity.

Multimodal and Knowledge-Based Assessment. In K-12 education, student responses often extend beyond text, incorporating drawings, diagrams, and equations, particularly in mathematics and science [States, 2013; Council, 2012]. Traditional text-based grading models struggle to interpret such multimodal inputs, limiting their ability to assess student understanding comprehensively [Li *et al.*, 2023b]. To address this, researchers have explored multimodal approaches in automated assessment. Models like CLIP [Radford *et al.*, 2021] enable joint encoding of hand-drawn images and text, improving grading accuracy, especially for responses requiring diagrams or equations [Baral *et al.*, 2021]. Additionally, knowledge graphs, such as K12EduKG [Chen *et al.*, 2018], provide structured representations to enhance assessment. Given the importance of multimodal reasoning in early education, our approach integrates textual, visual, and knowledge graph features for a more holistic evaluation of student understanding.

Human Bias in Grading. Bias in human grading is well-documented, with research highlighting systematic disparities where certain groups receive different scores despite similar performance [Li *et al.*, 2024; Gichoya *et al.*, 2023]. Factors like handwriting neatness and grader subjectivity contribute to unintended score variations [Li *et al.*, 2023b]. AI grading models may inherit such biases from human raters, necessitating debiasing strategies. Prior work has proposed mitigating grading bias through blinding demographic cues and adjusting scoring rubrics [An *et al.*, 2020]. Commercial

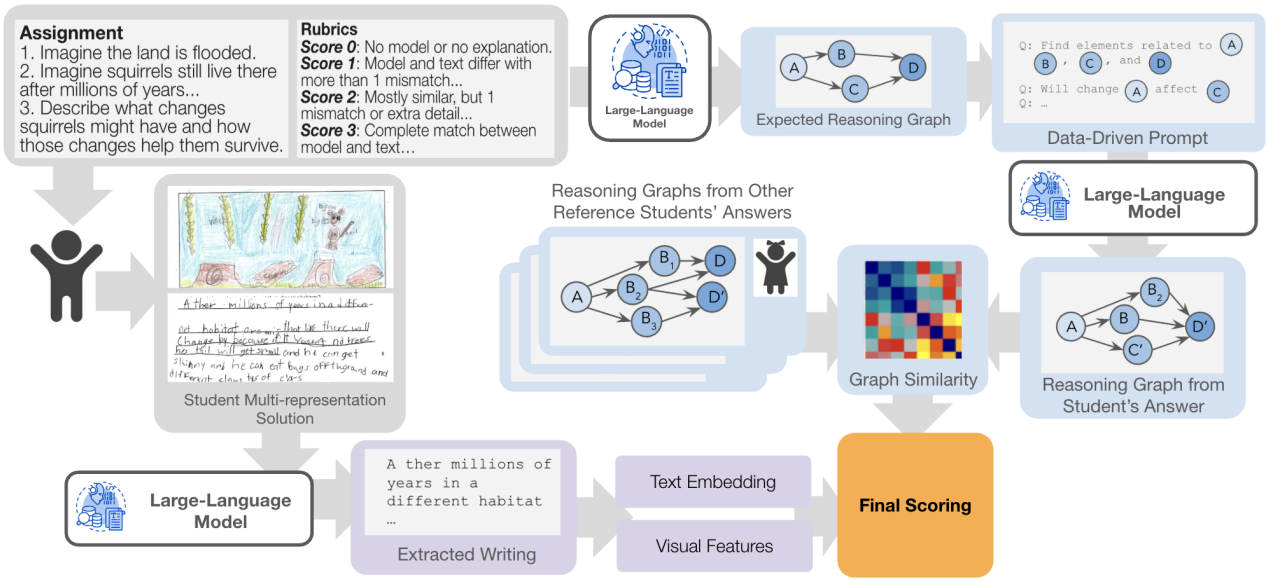

Figure 1: Overview of the automated grading framework. Given an open-ended question, students provide answers with handwritten text and drawings. We use LLMs to extract written text and generates a reasoning graph representing the conceptual structure. To assess alignment, the student's reasoning graph is compared against reference reasoning graphs constructed from other student submissions. Additionally, textual embeddings and visual features are incorporated into the final scoring module. By integrating causal reasoning, multimodal analysis, and structured assessment, this framework enhances the accuracy, fairness, and interpretability of automated grading in K-12 science education.

AI tools also emphasize bias reduction in grading [Schwartz *et al.*, 2022]. Our approach explicitly addresses presentation bias—how handwriting and visual clarity affect scoring—by introducing auxiliary metrics that enable the model to compensate for such biases. To our knowledge, this is the first work to integrate multimodal bias correction in an LLM-driven grading setting, particularly within the K-12 context.

3 Methodology

3.1 Problem Formulation

In open-ended questions, elementary students provide answers in the form of *a hand-drawn illustration* accompanied by *a handwritten explanation*. These multimodal responses require reasoning across both textual and visual modalities, making their assessment a complex task. To systematically evaluate the logical coherence of students' reasoning, we leverage multimodal Large Language Models to extract a *causal knowledge graph*, which encodes key concepts and causal relationships, thereby capturing the student's reasoning process. Additionally, we extract *visual and handwriting-based features* to account for presentation factors that may influence assessment. Figure 1 provides an overview of our workflow.

Formally, let T_i represent the transcribed text content from the answer of student i, D_i represent the student's drawing, H_i encode handwriting-related metadata (including *legibility, spelling accuracy, and writing style features*), and G_i denote the **causal knowledge graph**. We define the **true response quality** as:

$$y_i^{\text{true}} = f^*(T_i, D_i, H_i, G_i), \tag{1}$$

where f^* is the ideal, unbiased scoring function that accurately assesses the student's reasoning quality.

However, human-assigned scores often introduce biases, particularly due to subjective interpretation of *presentation factors* such as drawing clarity and handwriting quality. We denote the **human-assigned score** as:

$$y_i^{\text{human}} = y_i^{\text{true}} + B^*(D_i, H_i), \tag{2}$$

where $B^*(D_i, H_i)$ represents the bias introduced by non-reasoning factors such as drawing style or handwriting clarity.

Our objective is to develop an **automated scoring function** f that predicts a score $\hat{y}_i$ that closely approximates the **true response quality** y_i^{true}, while mitigating biases inherent in human grading. By integrating causal reasoning structures with multimodal features, our approach aims to provide a **fair, interpretable, and consistent** evaluation framework for assessing students' reasoning quality.

3.2 Human Bias Calibration

As described in Equation (2), human-assigned scores often incorporate systematic biases arising from *presentation factors*, such as penalties for poor handwriting or unclear diagrams. Since these factors do not directly reflect a student's reasoning ability, our goal is to estimate and correct for these biases to ensure a more objective assessment.

To quantify the extent of presentation-related bias, we introduce two **auxiliary presentation metrics**:

- **Handwriting Quality Score** $q_{\text{hand}}(H_i)$, which measures handwriting legibility based on features such as the number of misspelled words and a clarity score derived from character shape consistency.

- **Drawing Clarity Score** $q_{\text{draw}}(D_i)$, which assesses the informativeness and completeness of the student's diagram. This metric considers the number of key elements present in D_i and the overall level of visual detail.

To estimate the degree to which these presentation factors influence human-assigned scores, we fit a regression model:

$$y_i^{\text{human}} = \alpha_0 + \alpha_1 q_{\text{hand}}(H_i) + \alpha_2 q_{\text{draw}}(D_i) + \epsilon_i, \quad (3)$$

where α_1 and α_2 quantify the systematic effect of handwriting and drawing clarity on grading.

To mitigate these biases, we compute a **bias-corrected score** as follows:

$$y_i^{\text{true}} = y_i^{\text{human}} - (\alpha_1 q_{\text{hand}}(H_i) + \alpha_2 q_{\text{draw}}(D_i)). \quad (4)$$

These calibrated scores y_i^{true} serve as unbiased training targets for our automated scoring model, ensuring that it learns to evaluate responses based on their reasoning content rather than superficial presentation attributes.

3.3 Knowledge Graph Extraction and Alignment

A key component of our approach involves leveraging causal knowledge graphs to represent the semantics of student responses. We employ a large language model to extract a structured representation of the response in the form of a directed graph consisting of concepts and their relationships.

Positive Knowledge Alignment

To establish a structured evaluation of student responses, we first employ a *causal prompting* approach [Abdulaal *et al.*, 2024] to extract causal relationships from the assignment task and grading rubric. This process enables us to generate an *expected answer graph* $G^* = (V^*, E^*)$, where V^* represents key variables and E^* denotes their edges. To refine this representation, we further incorporate causal links identified in high-scoring student responses, ensuring that the expected graph reflects both expert knowledge and exemplary student reasoning patterns.

For each student submission, we prompt the LLM to extract a corresponding knowledge graph $G_i = (V_i, E_i)$ directly from the visual and textual inputs. Once both graphs are obtained, we compute a structural similarity score between G_i and G^* to assess conceptual alignment. Specifically, we define:

$$S_{KG}(i) = \frac{|G_i \cap G^*|}{|G^*|}, \quad (5)$$

where $S_{KG}(i)$ measures the proportion of expected causal relationships correctly captured in the student's response. This alignment score provides a robust, structured evaluation beyond simple text-matching approaches, allowing for a deeper assessment of student understanding based on the accuracy and completeness of their causal reasoning.

This knowledge graph alignment score is incorporated as a grading feature. Compared to conventional text matching or embedding-based similarity, structural graph alignment is more robust to paraphrasing and emphasizes the presence of correct relationships, making it particularly effective for evaluating student responses that involve free-text explanations or incomplete sentences.

Negative Knowledge Gap

To mitigate the risk of students inflating their scores by incorporating frequently occurring phrases from low-quality responses without demonstrating genuine reasoning, we introduce a majority-vote-based filtering mechanism. This approach identifies repetitive lexical patterns in low-scoring responses, detecting superficial answers that rely on rote memorization rather than substantive causal reasoning.

To quantify this phenomenon, we compute the cosine similarity between a student's response and a corpus of low-scoring responses. This helps distinguish between genuine logical reasoning and lexical redundancy. The cosine similarity is defined as:

$$S_{\text{gap}}(i) = \frac{\mathbf{T}_i \cdot \mathbf{L}^*}{\|\mathbf{T}_i\| \|\mathbf{L}^*\|}, \quad (6)$$

where $\mathbf{T}_i$ represents the vectorized form of student i's written answer, and $\mathbf{L}^*$ denotes the aggregated vector representation of historically low-scoring responses. If a response exhibits a high lexical similarity with these responses (exceeding a predefined threshold τ) while simultaneously achieving a low knowledge graph alignment score $S_{KG}(i) < \gamma$, it is classified as a **superficial response** and penalized accordingly.

This filtering mechanism ensures that assessments prioritize genuine reasoning over superficial textual similarity, reinforcing the importance of original causal explanations in student responses.

3.4 Multimodal Fusion

To predict student performance, we develop a **multimodal fusion model** that integrates information from four distinct modalities: (1) handwritten content, (2) student drawings, (3) handwriting quality, and (4) causal knowledge graph alignment. Given a student's response $\mathcal{R}_i = (T_i, D_i, H_i, G_i)$. We extract the following features:

Text Representation. A pre-trained transformer $\mathcal{E}_{\text{text}}$ extracts a dense representation of the transcribed handwritten answer:

$$\mathbf{f}_i^{(\text{text})} = \mathcal{E}_{\text{text}}(T_i) \in \mathbb{R}^{d_{\text{text}}}. \quad (7)$$

Drawing and Handwriting Representation. We also extract a structured feature vector from original inputs, combining visual semantics and drawing quality metrics:

$$\mathbf{f}_i^{(\text{draw})} = D_i \in \mathbb{R}^{d_{\text{draw}}}, \quad \mathbf{f}_i^{(\text{hand})} = H_i \in \mathbb{R}^{d_{\text{hand}}}. \quad (8)$$

Knowledge Graph Representation. Knowledge Graph alignment is captured through a similarity function:

$$\mathbf{f}_i^{(\text{KG})} = \mathcal{S}_{\text{KG}}(G_i, G^*) - S_{\text{gap}}(i) \in \mathbb{R}^{d_{\text{KG}}}, \quad (9)$$

where G^* represents the expected answer graph.

Multimodal Fusion and Prediction. The final representation is constructed via concatenate fusion:

$$\mathbf{z}_i = [\mathbf{f}_i^{(\text{text})} \| \mathbf{f}_i^{(\text{draw})} \| \mathbf{f}_i^{(\text{hand})} \| \mathbf{f}_i^{(\text{KG})}] \in \mathbb{R}^{d_{\text{fusion}}}. \quad (10)$$

Finally, a feed-forward neural network $F(\cdot; \Theta)$ predicts the final score:

$$\hat{y}_i = F(\mathbf{z}_i; \Theta), \quad (11)$$

Algorithm 1 LLM-Based Causal Knowledge Graph Extraction and Multimodal Scoring

Input: Student response $\mathcal{R}_i = (T_i, D_i, H_i, G_i)$
Output: Predicted score $\hat{y}_i$

Step 1: Feature Extraction
Extract text embeddings: $\mathbf{f}_i^{\text{text}} = \mathcal{E}_{\text{text}}(T_i)$
Extract drawing features: $\mathbf{f}_i^{\text{draw}} = D_i$
Extract handwriting features: $\mathbf{f}_i^{\text{hand}} = H_i$
Step 2: Causal Knowledge Graph Alignment
Construct student causal graph: G_i
Retrieve expert graph: G^*
Compute alignment score: $\mathcal{S}_{\text{KG}}(G_i, G^*)$
Compute gap score: $\mathcal{S}_{\text{gap}}(i)$
Compute final graph score: $\mathbf{f}_i^{\text{KG}} = \mathcal{S}_{\text{KG}}(i) - \mathcal{S}_{\text{gap}}(i)$
Step 3: Multimodal Fusion
Concatenate features:
$\mathbf{z}_i = [\mathbf{f}_i^{\text{text}} \| \mathbf{f}_i^{\text{draw}} \| \mathbf{f}_i^{\text{hand}} \| \mathbf{f}_i^{\text{KG}}]$
Step 4: Model Training (Bias-Corrected Target)
Compute bias-corrected score: $y_i^{\text{true}} = y_i^{\text{human}} - B^*(D_i, H_i)$
Train model $F(\cdot; \Theta)$ using input–label pairs $(\mathbf{z}_i, y_i^{\text{true}})$
Step 5: Inference
Predict score: $\hat{y}_i = F(\mathbf{z}_i; \Theta)$
return $\hat{y}_i$

where Θ is trained by minimizing:

$$\mathcal{L}(\Theta) = \frac{1}{N} \sum\nolimits_{i=1}^{N} \ell(F(\mathbf{z}_i; \Theta), y_i). \tag{12}$$

The detailed steps of the multimodal fusion and scoring process are outlined in Algorithm 1. Each modality offers complementary evidence of student understanding. Textual data conveys sequential reasoning, drawings capture spatial-visual knowledge, and causal graphs provide structural representations of logical thought. Multimodal fusion reduces ambiguity—if one modality is unclear, another may provide clarity. Bias correction further improves robustness by removing spurious correlations. By explicitly incorporating handwriting quality as a feature, we ensure that the final model learns to separate content-based merit from presentation-based penalties. Ultimately, this framework integrates LLM-driven reasoning with discriminative multimodal fusion, enhancing grading accuracy and fairness.

4 Experiments

4.1 Datasets

To evaluate the effectiveness of our approach, we introduce two open-ended K-12 educational datasets. These datasets contain scanned student assignments, each consisting of *handwritten textual responses* and *hand-drawn diagrams*, along with teacher-assigned scores on an ordinal scale from 0 to 3. Together, the two datasets comprise a total of 1,427 student responses.

Beyond serving as a benchmark for automated grading, these datasets present unique challenges due to their varying degrees of difficulty and openness. **Dataset 1** consists of responses to a question with a well-defined expected answer structure, allowing for more straightforward evaluation. In contrast, **Dataset 2** includes responses to a more open-ended question, where students demonstrate a wider range of reasoning patterns and response styles. This variation enhances the generalizability of grading models and better reflects real-world applications in educational assessment.

Each student response is provided in two formats to support different evaluation paradigms:

Raw Data. The raw dataset consists of scanned images of student responses, preserving the original handwriting and drawings. This format is suitable for direct image-based processing and enables few-shot prompting with large language models (LLMs), facilitating open-ended assessment without requiring structured annotations.

Processed Data. To facilitate structured analysis and automated assessment, we construct an annotated version of the dataset with extracted multimodal features, including:

- Transcribed handwritten text for text-based analysis,
- Automatically extracted causal knowledge graphs to capture reasoning structures,
- Identified key elements in student-drawn diagrams with corresponding semantic labels,
- Handwriting quality and drawing clarity scores assessing legibility and visual coherence.

The inclusion of both structured and unstructured responses in these datasets makes them well-suited for evaluating a range of machine learning models, including supervised learning, multimodal fusion, and large language model-based reasoning. The diversity in question complexity and response styles further enhances the applicability of these datasets to real-world educational assessment tasks, ensuring that automated grading methods can generalize across different levels of student reasoning and expression.

4.2 Baselines

To evaluate the effectiveness of our proposed approach, we compare it against a diverse set of baseline models, including both traditional image-based classification methods and large language model (LLM)-based grading approaches. These baselines provide insights into the impact of multimodal integration and structured knowledge representations on automated student response assessment.

Supervised Learning. We implement standard supervised learning models, including logistic regression, decision trees, and ResNet, which use raw image representations of student responses to predict scores. These models rely solely on visual information and serve as benchmarks for evaluating the efficacy of multimodal integration. As they do not incorporate textual reasoning or structured knowledge representations, their performance is expected to be limited, particularly for responses requiring conceptual understanding.

Multimodal Large Language Models. To evaluate the reasoning capabilities of LLM-based assessment methods,

Type	Model	Dataset 1			Dataset 2		
		Accuracy	Precision	F1	Accuracy	Precision	F1
Classification	Logistic Regression	76.54	75.67	46.06	33.89	34.25	33.87
	Decision Tree	69.27	35.09	30.90	31.54	29.41	31.29
	ResNet	70.39	74.18	75.34	37.85	36.45	36.18
Multimodal LLMs	ChatGPT-4o	61.03	76.04	66.64	42.13	39.85	41.22
	LLaMA 3.2-Vision-11B	60.42	71.69	60.42	34.83	36.48	34.62
	Qwen2.5-VL-7B	62.43	77.81	66.87	30.33	28.77	36.92
	DeepSeek-Janus-Pro-7B	47.85	67.12	54.95	27.63	30.06	20.37
Proposed Multimodal	LLaMA 3.2 Multimodal KG	74.63	77.59	76.08	41.23	42.38	44.13
	GPT-4o Multimodal KG	**81.94**	**68.36**	**74.54**	**61.22**	**61.13**	**61.06**

Table 1: Performance Comparison of Different Models

we employ state-of-the-art multimodal large language models, including GPT-4o, LLaMA3.2-Vision-11B, Qwen2.5-VL-7B, and DeepSeek-Janus-Pro-7B. These models process both handwritten text and visual content using CoT few-shot prompting to generate scores based on in-context learning [Cohn *et al.*, 2024]. While multimodal LLMs exhibit strong reasoning abilities, they operate in an implicit and unstructured manner, lacking explicit causal knowledge extraction and structured alignment mechanisms. Consequently, their scoring may exhibit inconsistencies, particularly in open-ended reasoning tasks where deeper conceptual understanding is required.

4.3 Results and Analysis

Table 1 presents the comparative performance of different models across the two datasets. The results demonstrate the advantages of multimodal integration and structured reasoning in automated student response grading.

Performance of Traditional Classification Models. Traditional supervised learning models, such as logistic regression, decision trees, and ResNet perform well on Dataset 1, achieving accuracy scores of 76.54% , 69.27%, and 70.39%. This is primarily due to simplicity of the task, which asks students to describe the effects of deforestation on squirrels. Most responses consist of simple drawings of felled trees accompanied by short, negative phrases such as "die" or "death." These responses exhibit clear visual patterns with minimal semantic variability, allowing traditional image classification models to achieve competitive performance, in some cases even surpassing vision-language models.

However, their performance deteriorates significantly on Dataset 2, with accuracies dropping to 33.89%, 31.54%, and 37.85%. This dataset presents a more abstract, open-ended question: "If the land is flooded, what adaptations might squirrels develop if they continue to live there after a million years?" Unlike Dataset 1, this task requires students to engage in conceptual reasoning, leading to highly diverse responses that vary in both textual explanations and graphical representations. As traditional classifiers rely solely on low-level visual features, they fail to generalize effectively, underscoring the limitations of purely image-based approaches for evaluating complex reasoning.

Multimodal LLMs Performance. Multimodal large language models (Multimodal LLMs) exhibit varying performance across the two datasets, revealing both their strengths and limitations in assessing open-ended K-12 responses. On Dataset 1, GPT-4o achieves an accuracy of 61.03%, while other models, such as LLaMA 3.2-Vision-11B and Qwen2.5-VL-7B, perform similarly. This suggests that for simpler tasks with structured expected answers, different multimodal models perform comparably, as the reasoning required remains relatively straightforward.

In contrast, Dataset 2 poses a greater challenge, leading to a decline in performance across all Multimodal LLMs. Notably, GPT-4o achieves the highest accuracy at 42.13%, while other models perform considerably worse. This trend suggests that larger models with more advanced reasoning capabilities, such as GPT-4o, generalize better to diverse and abstract responses. However, the overall drop in accuracy highlights the difficulty of open-ended educational assessments for existing multimodal models, which lack explicit causal reasoning mechanisms and structured answer alignment, resulting in inconsistencies in scoring.

Effectiveness of Our Approach. Our proposed method, which integrates multimodal learning with causal knowledge graph extraction, achieves the highest accuracy on both datasets. GPT-4o Multimodal KG outperforms the strongest baseline by +19.5% on Dataset 1 and +23.4% on Dataset 2. LLaMA 3.2 Multimodal KG also shows substantial F1 gains, with improvements of 16.7% and 9.5% on the two datasets, respectively. These results underscore the effectiveness of incorporating structured causal representations in grading student responses. By modeling causal relationships and leveraging multimodal features, our approach delivers more robust, interpretable, and generalizable scoring.

Impact of Dataset Complexity. The consistent performance drop from Dataset 1 to Dataset 2 across all baselines reflects the increasing challenge posed by open-ended assessments. While traditional classifiers struggle due to their reliance on visual features, Multimodal LLMs exhibit limitations in reasoning and answer alignment. The superior performance of our approach across both datasets underscores its ability to generalize to diverse response structures, making it more applicable for real-world educational assessment.

Ablation	Acc (%)	AUC	Precision	F1
w/o Text Embedding	54.08	74.64	53.18	53.27
w/o Graph	51.02	66.56	46.89	48.82
w/o Drawing	56.12	74.92	56.17	55.78
w/o Writing	59.18	71.79	60.08	58.90
w/ Full Model	**61.22**	**75.41**	**61.13**	**61.06**

Table 2: Ablation Study of Our Proposed Method

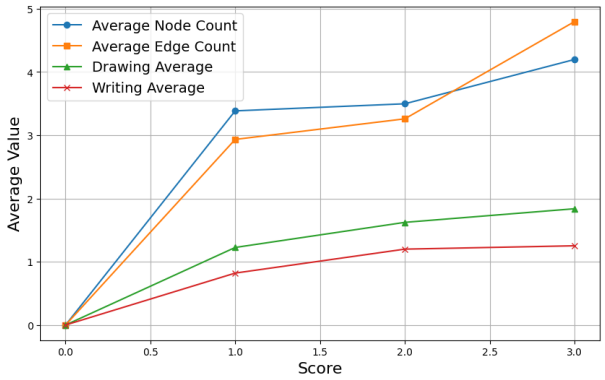

Figure 2: Graph Complexity and Modality Performance

4.4 Ablation Study

We report the results of our ablation study in Table 2, where we systematically removed different modalities to assess their individual contributions to model performance.

The most significant drop in accuracy occurs when removing the causal knowledge graph (**w/o Graph**), reducing accuracy from 61.22% to 51.02%, with a notable decrease in AUC (from 75.41% to 66.56%) and F1-score (from 61.06% to 48.82%). As shown in Figure 2, higher-scoring responses tend to have a greater number of nodes and edges, indicating that structured causal reasoning is crucial for capturing conceptual depth and distinguishing nuanced differences between responses.

Interestingly, removing the drawing modality (**w/o Drawing**) results in a smaller performance reduction (accuracy drops to 56.12%), suggesting that while student diagrams provide valuable complementary information, their impact is less critical compared to text and causal structures. Figure 2 further supports this finding, as drawing clarity shows a weaker correlation with score progression than knowledge graph features. Removing handwriting features (**w/o Writing**) leads to a moderate decline, with accuracy dropping to 59.18%. The figure suggests that handwriting quality slightly increases with score, but its contribution is relatively minor compared to other modalities. This result indicates that while handwriting legibility plays a role in presentation, it has a lesser impact on content-based assessment.

Overall, these findings reinforce the necessity of a multimodal approach, where causal reasoning, textual semantics, and visual elements collectively enhance student assessment. The full model achieves the highest accuracy (61.22%) and F1-score (61.06%), demonstrating the effectiveness of integrating multiple modalities for a comprehensive evaluation.

5 Future Work

While our framework demonstrates the effectiveness of multimodal grading, further improvements are necessary to enhance reliability, fairness, and scalability. One critical direction is strengthening the robustness of causal knowledge graphs, as LLM-generated graphs are prone to inconsistencies[Zečević *et al.*, 2023]. Integrating structured knowledge sources, such as educational ontologies or curated concept maps, could provide better grounding for causal relationships[Rousseau *et al.*, 2018]. Additionally, developing automated verification techniques can furthe help mitigate hallucinations, ensuring that extracted knowledge structures align with domain expertise.

Another key challenge is mitigating biases in automated grading. Although handwriting and drawing quality are incorporated as auxiliary features, presentation factors may still subtly affect scoring. Future research can explore fairness-aware training strategies that explicitly disentangle content-based evaluation from surface-level biases [Kamiran and Calders, 2012]. Incorporating human-in-the-loop mechanisms, where educators review and refine intermediate representations such as causal knowledge graphs, could further enhance transparency and trust in AI-assisted grading.

6 Conclusion

We propose a multimodal framework for automated grading in K-12 education, integrating causal knowledge graphs, handwriting recognition, drawing analysis, and text embeddings to enhance assessment accuracy and fairness. By leveraging LLMs for structured causal representation, our approach moves beyond surface-level text and visual similarity, capturing deeper reasoning structures critical for evaluating student understanding.

Experimental results demonstrate that multimodal integration improves grading reliability, with causal knowledge graphs playing a central role in differentiating reasoning quality. Ablation studies further confirm their impact, showing that structured representations significantly enhance grading accuracy. To support further research in AI-driven educational assessment, we introduce two multimodal K-12 science datasets, including transcribed handwritten responses, extracted knowledge graphs, and annotated visual elements.

By combining causal reasoning with multimodal analysis, our framework advances AI-assisted grading, making it more interpretable, fair, and applicable to real-world education.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant IIS-2212174. This study draws on data collected as part of the Multiple Literacies in Project-Based Learning project, which was funded by a division of the George Lucas Educational Foundation (APP13987). The findings and interpretations presented here are solely those of the authors and do not necessarily reflect the views of the Foundation. Data from this study are available from the corresponding author upon reasonable request.

References

[Abdulaal *et al.*, 2024] Ahmed Abdulaal, Adamos Hadjivasiliou, Nina Montaña-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C. Castro, and Daniel C. Alexander. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.

[Abu-Rasheed *et al.*, 2025] Hasan Abu-Rasheed, Constance Jumbo, Rashed Al Amin, Christian Weber, Veit Wiese, Roman Obermaisser, and Madjid Fathi. LLM-assisted knowledge graph completion for curriculum and domain modelling in personalized higher education recommendations. In *Proceedings of the IEEE Global Engineering Education Conference (EDUCON 2025)*, 2025.

[An *et al.*, 2020] Pengcheng An, Kenneth Holstein, Bernice d'Anjou, Berry Eggen, and Saskia Bakker. The ta framework: Designing real-time teaching augmentation for k-12 classrooms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2020.

[Baral *et al.*, 2021] Sami Baral, Anthony F Botelho, John A Erickson, Priyanka Benachamardi, and Neil T Heffernan. Improving automated scoring of student open responses in mathematics. *International Educational Data Mining Society*, 2021.

[Bonthu *et al.*, 2021] Sandeep Bonthu, S. Rama Sree, and M. H. M. Krishna Prasad. Automated short answer grading using deep learning: A survey. In *Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 61–78, 2021.

[Botelho *et al.*, 2023] Anthony Botelho, Sami Baral, John A Erickson, Priyanka Benachamardi, and Neil T Heffernan. Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. *Journal of computer assisted learning*, 39(3):823–840, 2023.

[Bui *et al.*, 2024] Tuan Bui, Oanh Tran, Phuong Nguyen, Bao Ho, Long Nguyen, Thang Bui, and Tho Quan. Cross-data knowledge graph construction for LLM-enabled educational question-answering system: A case study at HCMUT. In *Proceedings of the 1st ACM Workshop on AI-powered Question & Answering Systems*, 2024.

[Burstein *et al.*, 2013] Jill Burstein, Joel Tetreault, and Nitin Madnani. The e-rater® automated essay scoring system. In *Handbook of automated essay evaluation*, pages 55–67. Routledge, 2013.

[Caraeni *et al.*, 2025] Adriana Caraeni, Alexander Scarlatos, and Andrew Lan. Evaluating GPT-4 at grading handwritten solutions in math exams. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 2025.

[Chen *et al.*, 2018] Penghe Chen, Yu Lu, Vincent W. Zheng, Xiyang Chen, and Xiaoqing Li. An automatic knowledge graph construction system for K-12 education. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, pages 1–4, 2018.

[Chia *et al.*, 2024] Yew Ken Chia, Vernon Toh Yan Han, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16257–16266, 2024.

[Cohn *et al.*, 2024] Clayton Cohn, Nicole Hutchins, Tuan Le, and Gautam Biswas. A chain-of-thought prompting approach with llms for evaluating students' formative assessment responses in science. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23182–23190, 2024.

[Council, 2012] National Research Council. *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academy of Sciences, 2012.

[Gichoya *et al.*, 2023] Judy W. Gichoya, Kenol Thomas, Leo Anthony Celi, Nabile Safdar, Imon Banerjee, John D. Banja, and Saptarshi Purkayastha. Ai pitfalls and what not to do: mitigating bias in ai. *The British Journal of Radiology*, 96(1150):20230023, 2023.

[He *et al.*, 2023] P. He, I. C. Chen, I. Touitou, K. Bartz, B. Schneider, and J. Krajcik. Predicting student science achievement using post-unit assessment performances in a coherent high school chemistry project-based learning system. *Journal of Research in Science Teaching*, 60(4):724–760, 2023.

[Jiang and Bosch, 2024] Lan Jiang and Nigel Bosch. Short answer scoring with GPT-4. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, 2024.

[Kamiran and Calders, 2012] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[Khatibi *et al.*, 2024] Elahe Khatibi, Mahyar Abbasian, Zhongqi Yang, Iman Azimi, and Amir M. Rahmani. ALCM: Autonomous LLM-augmented causal discovery framework. *arXiv preprint*, arXiv:2405.01744, 2024.

[Kress, 2009] Gunther Kress. *Multimodality: A social semiotic approach to contemporary communication*. Routledge, 2009.

[Li *et al.*, 2023a] T. Li, I. C. Chen, E. Adah Miller, C. Miller, B. Schneider, and J. Krajcik. The relationships between elementary students' knowledge-in-use performance and their science achievement. *Journal of Research in Science Teaching*, pages 1–6, 2023.

[Li *et al.*, 2023b] T. Li, F. Liu, and J. Krajcik. Automatically assess elementary students' hand-drawn scientific models using deep learning of artificial intelligence. In *Proceedings of the 17th International Conference of the Learning Sciences-ICLS 2023*, pages 1813–1814. International Society of the Learning Sciences, 2023.

[Li *et al.*, 2023c] T. Li, E. Adah Miller, M. C. Simani, and J. Krajcik. Adapting scientific modeling practice for promoting elementary students' productive disciplinary engagement. *International Journal of Science Education*, 2023.

[Li *et al.*, 2023d] T. Li, E. Reigh, P. He, and E. A. Miller. Can we and should we use artificial intelligence for formative assessment in science. *Journal of Research in Science Teaching*, 60(6):1385–1389, 2023.

[Li *et al.*, 2024] T. Li, E. A. Miller, and P. He. Culturally and linguistically "blind" or biased? challenges for ai assessment of models with multiple language students. In *Proceedings of the 18th International Conference of the Learning Sciences-ICLS 2024*, pages 1323–1326. International Society of the Learning Sciences, 2024.

[Li, 2024] T. Li. *Cognitive Synergy: Exploring the Transformative Intersection of Human Intelligence and Artificial Intelligence in Designing Equitable Next Generation Science Assessments*. PhD thesis, Michigan State University, 2024.

[Lu *et al.*, 2023] Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. LLMScore: Unveiling the power of large language models in text-to-image synthesis evaluation. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[Ouyang *et al.*, 2022] Xiaomin Ouyang, Xian Shuai, Jiayu Zhou, Ivy Wang Shi, Zhiyuan Xie, Guoliang Xing, and Jianwei Huang. Cosmo: Contrastive fusion learning with small data for multimodal human activity recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing and Networking (MobiCom '22)*, pages 324–337. ACM, 2022.

[Ouyang *et al.*, 2023] Xiaomin Ouyang, Zhiyuan Xie, Heming Fu, Sitong Cheng, Li Pan, Neiwen Ling, Guoliang Xing, Jiayu Zhou, and Jianwei Huang. Harmony: Heterogeneous multi-modal federated learning through disentangled model training. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '23)*, pages 530–543. ACM, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021.

[Rousseau *et al.*, 2018] David Rousseau, Julie Billingham, and Javier Calvo-Amodio. Systemic semantics: A systems approach to building ontologies and concept maps. *Systems*, 6(3):32, 2018.

[Schwartz *et al.*, 2022] Renee Schwartz, Robert Schwartz, Apostol Vassilev, Katherine Greene, Lance Perine, Andrew Burt, and Patrick Hall. Towards a standard for identifying and managing bias in artificial intelligence. Technical report, US Department of Commerce, National Institute of Standards and Technology, 2022.

[Schwarz *et al.*, 2017] Christina V. Schwarz, Cynthia Passmore, and Brian J. Reiser. *Helping students make sense of the world using next generation science and engineering practices*. National Science Teachers Association, 2017.

[States, 2013] NGSS Lead States. *Next generation science standards: For states, by states*. National Academies Press, 2013.

[Sultan *et al.*, 2016] Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. Fast and easy short answer grading with high accuracy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1070–1075, 2016.

[Sung *et al.*, 2019] Chin-Yew Sung, Tanmay I. Dhamecha, and Nikhil Mukhi. Improving short answer grading using transformer-based pre-training. In *Proceedings of the 20th International Conference on Artificial Intelligence in Education*, pages 469–481, 2019.

[Wang *et al.*, 2019] Qi Wang, Claire Boudreau, Qixing Luo, Pang-Ning Tan, and Jiayu Zhou. Deep multi-view information bottleneck. In *Proceedings of the 2019 SIAM International Conference on Data Mining (SDM)*, pages 37–45. Society for Industrial and Applied Mathematics, 2019.

[Wang *et al.*, 2020] Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1828–1838. ACM, 2020.

[Zečević *et al.*, 2023] Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023.