Automated Diagnosis of Students' Number Line Strategies for Fractions

Zhizhi Wang
Rutgers University
zw393@cs.rutgers.edu

Dake Zhang
Rutgers University
dake.zhang@gse.rutgers.edu

Min Li
University of Washington
minli@u.washington.edu

Yuhan Tao
Columbia University
yt2905@tc.columbia.edu

Abstract

Fraction number lines are widely recognized as an effective tool for supporting students' conceptual understanding of rational numbers, yet their abstract nature makes them challenging for students to use and for teachers to evaluate reliably. This study introduces Visual Translator (VT), an AI-based tool designed to automatically diagnose students' problem-solving strategies and error types in handwritten fraction number line tasks. VT employs object detection models trained on a curated subset of 1,134 annotated student responses from the ASSISTments Mathnet dataset, and generates structured text descriptions of key visual elements, which can then be processed by large language models (LLMs) for higher-level reasoning. Preliminary results indicate that VT outperforms GPT-4o and Grok in fraction value identification (Precision = 0.611, Recall = 0.582), while achieving substantially higher accuracy in location detection (mAP@50 = 0.88) compared to Gemini-2.5 Pro (0.11). These findings demonstrate the promise of combining computer vision with generative AI to improve automated diagnosis of students' mathematical strategies.

1 Introduction

Fractions represent a critical transition in mathematics learning, serving as a bridge between whole numbers and rational numbers and laying the foundation for later understanding of proportional reasoning, algebra, and real numbers (Siegler et al., 2011; Siegler and Pyke, 2013). However, decades of research have shown that students frequently struggle with fractions, often treating them as two whole numbers rather than as magnitudes on a continuous scale (Ni and Zhou, 2005). To address these difficulties, instructional researchers have emphasized the use of number lines as a visual and conceptual tool for representing fractions. A number line highlights relative magnitude, density, and equivalence of rational numbers, making it particularly effective for developing conceptual understanding (Stewart et al., 2008). As such, number lines have been widely recommended in standards and curricula, including the Common Core State Standards for Mathematics (Association et al., 2010).

Despite their benefits, number lines also present challenges for students. Siegler and colleagues identified two primary strategies (Siegler et al., 2011) used by students: segmentation strategies, which involve partitioning the interval between 0 and 1, and numerical transformation strategies, which involve mapping fractions onto known reference points. Errors commonly observed include uneven segmentation, the use of incorrect units, or misapplied transformations (Bright et al., 1988; Lamon, 2007). These errors are not merely procedural slips; they reflect deeper misconceptions about the nature of rational numbers (Lamon, 2007). For teachers, especially novice teachers, diagnosing these misconceptions from handwritten number line representations is both cognitively demanding and time-consuming (Zhang et al., 2016). Consequently, there is a pressing need for scalable tools that can assist teachers in analyzing student responses and identifying error types with accuracy and consistency.

The rapid advancement of artificial intelligence (AI) provides new opportunities to address this need. Automated scoring systems have demonstrated success in domains such as essay grading, short-answer evaluation, and mathematical problem solving (Lockwood, 2014; Dikli, 2006). Recent work has also explored the use of computer vision and large language models (LLMs) to interpret drawn diagrams and models. For example, Lee and Zhai reported limited success in using GPT-4o to grade student-drawn science models, with accuracy ranging from 0.2 to 0.6, highlighting the challenges of reliably recognizing children's handwritten and

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 178–184

diagrammatic representations (Lee and Zhai, 2023). Similarly, early experiments with GPT-4o and related multimodal models suggest that, while LLMs excel in natural language reasoning, their image-processing capabilities remain insufficient for fine-grained educational diagnostics such as interpreting number lines.

In this study, we evaluate VT against both human-annotated ground truth and state-of-the-art LLMs (GPT-4o, o3, Gemini-2.5, Grok) on two key tasks: (a) recognizing handwritten fraction values, and (b) detecting the locations of visual elements on number lines. Our preliminary results show that VT achieves substantially higher accuracy in location identification (mAP@50 = 0.88 vs. 0.11 for Gemini-2.5 Pro) and competitive accuracy in fraction value recognition, outperforming GPT-4o and Grok. Beyond empirical results, our contributions are threefold:

- We manually labeled over one thousand student responses featuring fraction number lines, creating a domain-specific dataset with fine-grained annotations of key visual elements (ticks, endpoints, and handwritten values).
- We trained the Visual Translator (VT) on this dataset, tailoring it specifically for fraction number line tasks to detect and interpret key visual information from students' handwritten solutions.
- We designed metrics to assess the accuracy of models in capturing key visual information in students' work, and conducted extensive experiments comparing VT with leading multimodal LLMs, including GPT-4o, GPT-o3, Gemini-2.5 Pro, and Grok-2.

2 VT Model

2.1 Data Preparation

The first step in developing the VT model was to identify student responses containing fraction number lines. From the full MathNet dataset of 3.8 million images, we initially filtered 139,000 fraction-related items using keywords extracted from the associated JSON metadata. From this subset, we manually labeled a small number of images that clearly contained number lines to serve as seed data.

To expand the labeled dataset efficiently, we trained a YOLOv8 model on the seed images to detect number lines and applied it to additional candidate images. Predictions from the model were then manually verified to confirm their relevance. This iterative process, in which the model guides the selection of images for human annotation, effectively implements an active learning strategy, concentrating labeling effort on the most informative samples and improving data collection efficiency.

Table 1: Statistics of Key Elements.

Key Elements Type	Number of Instances
Fraction	8199
Tick	8385
0	2474
1	4447
2	2236
3	2054
4	2182
5	1732
6	2387
7	606
8	1204
9	476

After identifying 1,134 confirmed images featuring 0–1 fraction number lines, we conducted fine-grained annotations of key visual elements, which are identified by our educational experts, including tick marks, digits(0-9), and fractions. Finally, the dataset contains more than 8,000 fraction labels, over 8,300 tick marks, and thousands of digit labels (0–9). Detailed statistics of the labeled dataset are summarized in Table 1. All annotations were completed by our graduate assistants using Roboflow[1], a comprehensive platform for data annotation, model training, and deployment. Each key element was enclosed within a bounding box of a distinct color and assigned a unique label, as shown in Figure 1. The platform allows export of labeled information into various formats, including .txt, .json and other supported formats, enabling users to directly download the annotation files. These annotations serve as the foundation for subsequent model training, evaluation, and automated diagnostic tasks.

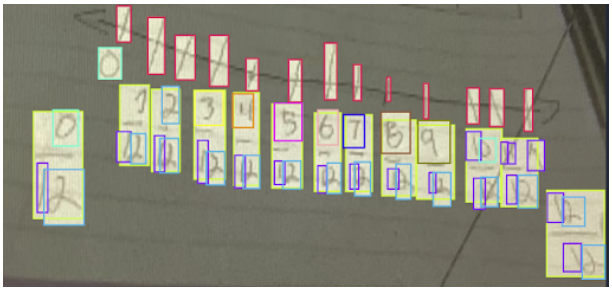


Figure 1: An example for labeling work.

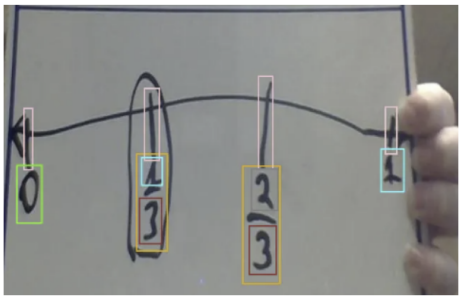
[1] https://roboflow.com

2.2 Textual Description Generation

After preparing and annotating the dataset, the VT model was trained to generate textual descriptions of a student's work from a given input image. For each image, the model produces a description that includes the key visual elements, their corresponding labels, and spatial coordinates. Additionally, the model identifies whether the number line is a 0–1 number line by detecting the presence of leading 0 and ending 1.

When a key element corresponds to a fraction, the model further derives the fraction value from the detected digits in the bounding box of a fraction. Specifically, a post-processing clustering algorithm groups digits associated with each fraction into two sets, corresponding to the numerator and denominator. The final fraction value is then computed from these clusters, ensuring accurate reconstruction of the student's intended fraction. Besides that, for a detected fraction, it is linked to its corresponding tick (if present) in the form of "F0-T1", where F0 denotes the first fraction and T1 denotes the second tick. The indices of fractions (i.e., F0, F1, F2 ...) and ticks (T0, T1, T2 ...) were automatically generated based on the left-to-right order of the top-left coordinates of their bounding boxes.

Overall, this process provides a structured textual summary of the detected key elements in a student's response, including their spatial information and numerical content. These textual summaries serve as standardized inputs for downstream diagnostic tasks, enabling automated error analysis and strategy classification. An example for textual generation is shown in Figure 2.



The key elements are interpreted via visual translator. Their coordinates are represented as outlined boxes (top-left, bottom-right).

There is a zero on the left side of the number line. Its coordinate is ((18.51, 451.54), (54.55, 533.06))

There is a one on the right side of the number line. Its coordinate is ((529.37, 428.74), (562.69, 489.58))

There are 4 ticks. Their coordinates are: ((28.04, 371.14), (41.75, 458.93)), ((197.60, 343.85), (219.68, 445.09)), ((357.89, 326.44), (378.36, 457.85)), ((535.95, 369.70), (552.90, 436.53)),

There are 2 fractions. Their coordinates are: ((186.31, 435.92), (230.59, 572.45)), ((335.48, 452.32), (388.54, 618.10)), The fraction numbers from left to right are: ['1/3', '2/3']. 1st fraction is associated with 2nd tick. 2nd fraction is associated with 3rd tick.

Figure 2: A demo example for textual generation.

2.3 Model Development

We developed the VT model in the following pipeline:

2.3.1 Key Element Detection

We trained an object detection model from the YOLO series to identify ticks, digits, and fractions. Training was conducted on the Roboflow platform, which provides resources optimized for small object detection in real-time. The labeled dataset enabled the model to learn the visual appearance and spatial layout of key elements. Our best-performing model achieved a mean Average Precision at IOU 0.5 (mAP@50) of 0.88 on the validation set, demonstrating high accuracy in detecting fine-grained handwritten components.

2.3.2 Model Deployment

The trained VT model is deployed to detect key elements—ticks, digits, and fractions—while returning their corresponding labels and spatial coordinates. The deployment is hosted on Roboflow and accessible via an API, which allows external systems to directly query the model. This design enables smooth integration into various downstream applications without requiring local installation or complex setup.

2.3.3 Web-based Interface

To further enhance usability, we developed a web application hosted on Hugging Face Spaces, offering an interactive interface for educators and researchers. Through this platform, users can:

- Upload an image of a student's work.
- Visualize detection results superimposed on the original image.
- Automatically reconstruct fraction values by clustering detected digits into numerators and denominators.
- Generate a textual summary of all identified key elements along with their coordinates.
- Download the complete results as a JSON file for integration into other pipelines.

The web service is hosted on Hugging Face Spaces and can be accessed at MathNet VT Model Web Platform. Access is granted via the invitation token **RU_MATHNET_VT**.

3 Evaluation

Unlike conventional object detection models, the VT system is designed not only to detect visual elements but also to generate structured textual descriptions that capture the key information in students' handwritten number line tasks. While standard detection metrics such as mAP@50 provide useful references, they are insufficient to directly reflect performance on our educational task, where the ultimate goal is to recover meaningful mathematical content (e.g., fractions, their values, and their associations with ticks).

To address this gap, we manually annotated additional ground truth information–including fraction values ordered from left to right according to the bounding box locations of fractions, as well as the associations between fractions and their corresponding ticks–and developed a set of tailored evaluation metrics that complement traditional object detection measures. These metrics assess (i) the accuracy of recognized fraction values and (ii) the correctness of their associations with ticks. Finally, we designed a weighted composite score that integrates our custom metrics with conventional detection metrics. This composite score serves as a more intuitive and holistic evaluation of the performance of VT, balancing both visual detection accuracy and the recovery of meaningful mathematical content from student work.

3.1 Additional Ground Truth Construction

In addition to the bounding box annotations used for model training, two further steps were conducted to establish reliable ground truth:

- **Fraction Value Annotation.** Expert graduate assistants manually read each student's handwritten work and labeled fraction values (e.g., 1/8, 2/8, 3/8). These values were aligned with bounding boxes to create a verified mapping between visual regions and fraction numbers.
- **Fraction–Tick Association.** Fractions were linked to their corresponding ticks (if present) using index pairs (e.g., F0–T1), where indices were assigned based on the left-to-right order of their bounding boxes.

As illustrated in Figure 3, this multi-level annotation ensured that ground truth covered both fraction recognition and structural relationships in the number line.

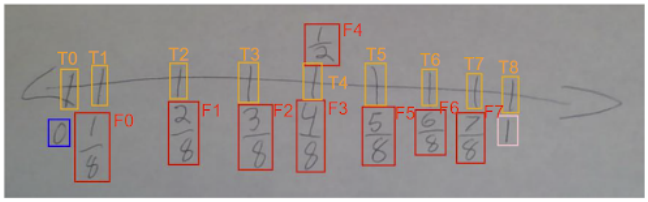


Figure 3: Additional ground truth annotation.

3.2 Metrics for Fraction Value Identification

To evaluate fraction recognition, we designed two complementary accuracy indices:

- **Jac Index (Order-independent).** Measures the set overlap between predicted and ground-truth fraction values, ignoring their order. It reflects the completeness of detection.
- **Seq Index (Order-sensitive).** Measures the length of the longest subsequence of correctly predicted fractions that also appear in the correct order. This is stricter than the Jac Index, as it penalizes out-of-order predictions.

In addition, we computed precision and recall for fraction values:

$$\text{Precision} = \frac{\#\text{ of correctly detected fraction values}}{\#\text{ of all detected fraction values}}$$

$$\text{Recall} = \frac{\#\text{ of correctly detected fraction values}}{\#\text{ of all ground-truth fraction values}}$$

For example, assume the ground truth fractions are 0/3, 1/3, 2/3, 3/3, If the model predicts 0/3, 3/3, 2/3, then three of the predicted fractions are correct under the jac index metric. In this case, the model achieves a precision of 1.0 (since three out of three predictions are correct) and a recall of 0.75 (since it misses one ground-truth fraction, 1/3). While under the seq index, its precision is only 0.67 because 2/3 is out of the right order and its recall is 0.5 (since it misses two ground-truth fractions, 1/3 and 2/3).

3.3 Metrics for Relationships between Fractions and Ticks

In addition to evaluating individual fraction values and key element locations, a crucial aspect of analyzing students' number line work is capturing the spatial and logical relationships between fractions and their corresponding ticks. Correctly identifying these relationships ensures that each fraction

is accurately mapped to its intended position on the number line, which is essential for subsequent diagnostic analyses.

To assess this, we introduce relationship-specific metrics that compare the associations generated by VT with ground-truth annotations. Each fraction in the ground truth is linked to a specific tick (if present) using index pairs (e.g., F0–T1), where the indices are determined based on the left-to-right ordering of the top-left coordinates of their bounding boxes. VT's predicted fraction–tick pairs are then matched against these ground-truth pairs. Accuracy is calculated as the proportion of correctly identified fraction–tick relationships over all annotated pairs. In addition, we compute precision and recall for these relationships to provide a more detailed assessment of VT's performance in capturing fraction–tick associations The overall fraction–tick score is summarized by its F1-score, which we will introduce later.

3.4 Composite Score

To provide a single, interpretable measure of VT's overall performance, we designed a composite score that integrates both elemental and relational information extracted from student work. To account for both precision and recall, we adopt the F1 score as a comprehensive performance metric, which is computed as follows:

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The textual description score is calculated as a weighted average of these per-element F1 scores:

$$\text{Textual Description Score} = \sum_i 100 * F1_i * W_i$$

where i indexes the key element categories and W_i denotes their user-defined weights and $\sum_i W_i = 1$. This formulation provides flexibility, as users can assign greater importance to specific elements (e.g., fractions over digits) depending on instructional or diagnostic needs. In addition, detailed per-element results (precision, recall, and F1) are retained to facilitate fine-grained analysis.

To achieve a comprehensive assessment, we integrate the textual description score with the accuracy of fraction–tick relationship detection. Specifically, the final composite score is computed as:

$$\begin{aligned}\text{Composite Score} = {} & W_{ft} * \text{Fraction-Tick Score} \\ & + W_{ts} * \text{Textual Description Score}\end{aligned}$$

where W_{ts} is the weight ratio of textual description score, W_{ft} is the weight ratio of fraction-tick score and $W_{ts} + W_{ft} = 1$. Since fraction–tick relationships depend on the accurate detection of both fractions and ticks, missing either element directly leads to a missing relationship. Therefore, it is advisable to assign a much higher weight W_{ts} than to W_{ft}, e.g. $W_{ts} = 0.8$ and $W_{ft} = 0.2$. By combining elemental accuracy and relational correctness into a single metric, this composite score provides an interpretable and holistic evaluation of VT's effectiveness in extracting both low-level visual details and higher-order relational information from student work.

4 Experiments

We conducted a series of experiments to evaluate the effectiveness of VT in comparison with state-of-the-art large language models (LLMs). Specifically, we first benchmark VT against Grok-2, Gemini-Pro 2.5, GPT-4o, and GPT-o3 on the task of fraction value detection, where accuracy serves as the evaluation metric. In addition, we assess the composite score of VT and Gemini-Pro 2.5, as Gemini-Pro 2.5 is the only publicly available LLM known to support image segmentation, i.e., the ability to process element-level bounding box coordinates.

4.1 Accuracy on Fraction Values Detection

To ensure fairness, the evaluation was conducted on a set of 227 images sampled from both the validation and test datasets, none of which were used to train VT. All models were tested under the same experimental settings.

The results are shown in Table 2. Gemini-2.5 Pro achieved the highest precision and recall across both the Jaccard index and sequential index metrics, demonstrating its strong capability in fraction detection. However, VT consistently outperformed GPT-4o and Grok-2,though its performance was slightly lower than that of GPT-o3. Overall, all models performed somewhat worse on the sequential index than on the Jaccard index, suggesting that capturing the correct ordering of elements remains a challenging aspect.

4.2 Comprehensive Evaluation

The comprehensive evaluation was performed on a combined set of 177 images drawn from our validation and test data. We evaluated both models using our proposed composite score metric, with

Table 2: Accuracy results on fraction values detection.

Model	precision (jac index)	recall (jac index)	precision (seq index)	recall (seq index)
Grok-2	0.320	0.387	0.243	0.293
GPT-4o	0.459	0.408	0.357	0.332
Gemini-2.5-pro	**0.646**	**0.668**	**0.520**	**0.538**
VT	0.560	0.499	0.446	0.403
GPT-o3	0.594	0.527	0.490	0.446

Table 3: Composite scores of VT and Gemini-2.5Pro

Model	VT	Gemini-2.5Pro
Composite score	**66.8**	15.8
Textual description score	**73.0**	18.1
Fraction-Tick score	**41.9**	6.2

the detailed results presented in Table 3, including the composite score along with textual description score and fraction-tick score. The Textual description score itself is a weighted average reflecting the accurate detection of key elements, with weights defined as 'fraction': 0.5, 'tick': 0.4, 'one': 0.05, 'zero': 0.05. The final Composite score is then calculated by combining the Textual description score (representing key element detection, 'ke') and the Fraction-Tick score (representing relationship between a fraction and its corresponding tick, 'tick2frac') with weights of $W_{ke} = 0.8$ and $W_{tick2frac} = 0.2$, respectively.

As illustrated in Table 3, our VT model demonstrates a commanding lead across all metrics. VT achieved a composite score of 66.8, which is more than four times higher than the 15.8 scored by Gemini-2.5Pro. This significant gap is consistent across the sub-metrics: VT scored 73.0 on textual description and 41.9 on fraction-tick relationships, compared to Gemini-2.5Pro's scores of 18.1 and 6.2.

The stark performance disparity underscores the critical importance of domain-specific training for specialized, high-precision tasks. While LLMs like Gemini-2.5Pro possess extensive general knowledge, they struggle to accurately parse the fine-grained, structured information required by our task without targeted fine-tuning. This outcome strongly indicates that Large Language Models (LLMs) do not serve as an infallible "oracle" or a universal solution for all problems.

5 Conclusion

In this paper, we presented VT, a specialized vision-language model designed to parse key semantic information from student-produced diagrams of fraction number lines. Departing from conventional object detection methods that focus primarily on localization, VT generates a structured textual representation that encapsulates not only elemental components (e.g., digits, ticks, fractions) but also their crucial relational associations. To facilitate a rigorous and fair evaluation, we have contributed a manually annotated dataset of over 1,000 student drawings and proposed a suite of tailored metrics, including accuracy on fraction values, fraction-tick relationship metrics, and a weighted composite score that provides a comprehensive assessment of model performance. Our empirical results demonstrate that VT significantly outperforms general-purpose Large Language Models. This finding suggests that while LLMs offer broad capabilities, they are not a universal solution; for domain-specific tasks requiring fine-grained interpretation of private data, developing and training specialized models remains a necessary and effective approach for robust information extraction.

References

National Governors Association and 1 others. 2010. Common core state standards. *Washington, DC*.

George W Bright, Merlyn J Behr, Thomas R Post, and Ipke Wachsmuth. 1988. Identifying fractions on number lines. *Journal for research in mathematics education*, 19(3):215–232.

Semire Dikli. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1).

Susan J Lamon. 2007. Rational numbers and proportional reasoning: Toward a theoretical framework for research. *Second handbook of research on mathematics teaching and learning*, 1(1):629–668.

Gyeong-Geon Lee and Xiaoming Zhai. 2023. Nerif: Gpt-4v for automatic scoring of drawn models. *arXiv preprint arXiv:2311.12990*.

Jane Lockwood. 2014. Handbook of automated essay evaluation current applications and new directions mark d. shermis and jill burstein (eds.)(2013). *Writing & Pedagogy*, 6(2):437–442.

Yujing Ni and Yong-Di Zhou. 2005. Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. *Educational psychologist*, 40(1):27–52.

Robert S Siegler and Aryn A Pyke. 2013. Developmental and individual differences in understanding of fractions. *Developmental psychology*, 49(10):1994.

Robert S Siegler, Clarissa A Thompson, and Michael Schneider. 2011. An integrated theory of whole number and fractions development. *Cognitive psychology*, 62(4):273–296.

James Stewart, Lothar Redlin, and Saleem Watson. 2008. *Precalculus: Mathematics for calculus*, 5 edition. Brooks/Cole.

Dake Zhang, Pamela Stecker, Sloan Huckabee, and Rhonda Miller. 2016. Strategic development for middle school students struggling with fractions: Assessment and intervention. *Journal of Learning Disabilities*, 49(5):515–531.