HOW VISION-LANGUAGE MODELS CAN SUPPORT EYE-TRACKING ANALYSIS IN MATHEMATICS EDUCATION: A METHODOLOGICAL STUDY ON NUMBER LINE ACTIVITIES

Parviz Asghari[1], Maike Schindler[2], Franziska Schulte[2] & Achim J. Lilienthal[1,3]

[1]Technical University of Munich, [2]University of Cologne,
[3]Örebro University

Eye tracking (ET) is a valuable tool for identifying students' learning processes, also in number line activities. While prior research has shown that AI can successfully identify student number line processes, the approaches operate as a "black box" and require extensive data collection and manual labelling. Vision-Language Models (VLMs) offer a promising, generally applicable, alternative. They can analyze ET data without prior labelling by leveraging domain knowledge. The aim of this study is to investigate how effectively VLMs can identify student number line processes. Based on an ET study with 136 fifth graders, we found that our VLM framework VETA (VLMs for ET Analysis) achieved very high agreement with the researchers' manual coding (Cohen's $\kappa \geq 0.80 - 1.0$), demonstrating reliability and potential for ET analyses.

INTRODUCTION

Understanding how students engage with mathematical tasks is crucial for enhancing teaching practices and promoting students' mathematical learning. Eye-tracking (ET) technology offers unique insights into students' processes for solving mathematical tasks by capturing gaze patterns, which reflect their cognitive processes (Schindler et al., 2019). Number lines are important representations in mathematics education. ET has been used, for example, for analyzing students' processes in number line activities. Simon and colleagues (2023) have identified various processes that students use to locate numbers on a marked number line. Gaze data can reveal that students use direct locating (from the beginning, middle, end), when their gaze goes directly to the correct position on the number line, or they use reference points (starting point, midpoint, end point) and continue counting from there (Simon et al., 2023). These insights are valuable for educators who want to support students according to their different needs. However, analysing ET data qualitatively is time-consuming and requires domain expertise. Typically, researchers have manually inspected gaze-overlaid recordings to label students' processes. It requires expertise in ET and in mathematics education to interpret which gaze patterns reflect specific solution processes.

Recent work started to consider automating ET analysis with machine-learning (AI) methods (e.g., Schindler et al., 2019). Some approaches rely on basic first-order statistics, e.g., fixation count, mean fixation duration, and time to first fixation, whereas others model gaze sequences and uncover group-level patterns in students (e.g., Simon

2026. In H. Krzywacki, M. Hähkiöniemi, E. Haataja, M.S. Hannula & P. Hästö (Eds.). *Proceedings of the 49th Conference of the International Group for the Psychology of Mathematics Education: Research Reports*, pp. 51-58. PME.

et al., 2023). Building on these developments, supervised machine-learning models have been used to identify and classify students' problem-solving processes from ET data (e.g., Boels et al., 2024). Despite their promise, supervised approaches face several limitations: they require large labelled datasets to capture the diversity of student processes, they often generalize poorly to new tasks and therefore demand repeated cycles of data collection and retraining, and they mainly operate as "black boxes" with limited explainability. This lack of transparency makes it difficult for educators to understand how decisions are made, reducing the pedagogical value of the results.

Current progress in foundation models, especially vision-language models (VLMs), offers a promising way to analyze ET data. A VLM is an advanced AI model that can understand and reason about images and text at the same time, so it can describe what it sees and answer questions in words (Bordes et al., 2024). Because of this, VLMs can work with multimodal input (images and text), use domain knowledge given in a prompt, and explain why they reach a certain conclusion.

Unlike traditional supervised machine-learning methods, VLMs can be used even when there is little or no labelled ET data by using expert pedagogical knowledge provided in the prompt. For example, in a number-line estimation task, a VLM can "look" at the task stimulus (the number line, endpoints, and marks) and an ET visualization (e.g., a scanpath, fixation maps, or heatmaps). Based on the expert information in the prompt, it can then analyze how the student approached the task.

To leverage the capabilities of VLMs and to address the limitations of prior machine learning models for ET analysis (need for extensive labelled datasets, limited performance when tasks change, and weak integration of domain knowledge), we developed a VLM-based pipeline, VLMs for ET Analysis (VETA), that is fed with participants' visually represented ET data (fixation maps; see Fig. 2) and then infers students' processes. Correspondingly, we pursue the research question: *How accurately can VLMs identify students' processes in number-line activities compared with manual labelling by researchers?*

LITERATURE REVIEW

Conceptual Framework

In this study, we follow the conceptualization of eye movements as indicators for attention processes (e.g., Duchowski, 2017). ET allows researchers to observe which pieces of information students focus on while working on mathematical tasks (e.g., Andrá et al., 2015), when an individual's gaze directly focuses on a stimulus, referred to as overt visual attention (Duchowski, 2017). However, since covert attention—the processing of stimuli in peripheral or extrafoveal vision—cannot be captured using ET (Klein & Ettinger, 2019), any interpretation of gaze data must take these inherent limitations of the method into account while also considering the specific mathematical content and the characteristics of the tasks being analyzed (Schindler et al., 2025).

Students' visual attention can reveal, for example, which elements of a number line they focus on when locating a number, and visual attention to different parts of the representation provides insight into students' orientation on the number line. Students' visual attention while working on these tasks can thus offer insights into their mathematical processes. However, interpreting students' visual attention into meaningful classifications of processes is a significant challenge. It requires domain expertise, and it is also a time-consuming process. These practical challenges have long limited the scalability of ET as an analytical tool in educational research.

AI use in Eye tracking Analysis

To use ET in applications such as adaptive learning systems and intelligent tutoring systems, automated analysis is needed. Supervised Machine Learning is a branch of AI where a model is trained on labelled data to predict outputs from inputs (Murphy, 2012), has been used to identify students' processes from gaze data or to distinguish between student groups. For example, Boels and colleagues (2024) used a random forest classifier to identify underlying student processes. Kennel and colleagues (2025) extracted gaze features from students' ET data while solving graphical differentiation problems and trained supervised machine learning classifiers (including K-nearest neighbours, Random Forest, and Support Vector Machines) to identify problem-solving difficulties in near real time. Recently, Hao and colleagues (2025) used ET glasses to collect Area of Interest-based fixation, saccade, and revisit features while students solved context-based problems, and applied partial least squares regression to predict math scores.

However, supervised machine learning models often face dataset challenges, such as labelling issues, imbalanced datasets, and learning with a low number of samples per class. Additionally, supervised machine learning often functions like a black box and does not provide a human-readable explanation for its decisions, which makes it difficult for educators or teachers to interpret and trust its output (Boels et al., 2024).

METHOD

The VETA pipeline consists of two main modules. An *information extraction module* and *process identification module (see Fig. 1)*. The input of the *information extraction module* is the fixation map (fixations are periods during ET in which the eye remains relatively stable on a specific location, fixation maps are spatio-temporal representations that aggregate ET fixations, e.g., showing longer durations as larger circles). It extracts the sequence of visited hatch marks on the number line task (e.g., 910→900→…→970). Then, the *identification module* attempts, based on this visited sequence, to identify the student's process (Fig. 1). Each module utilizes Gemini 2.5 Pro. We use mathematics education researchers' knowledge to define a prompt for hatch-mark detection and describe how the system should identify the processes from the sequence of detected hatch marks.

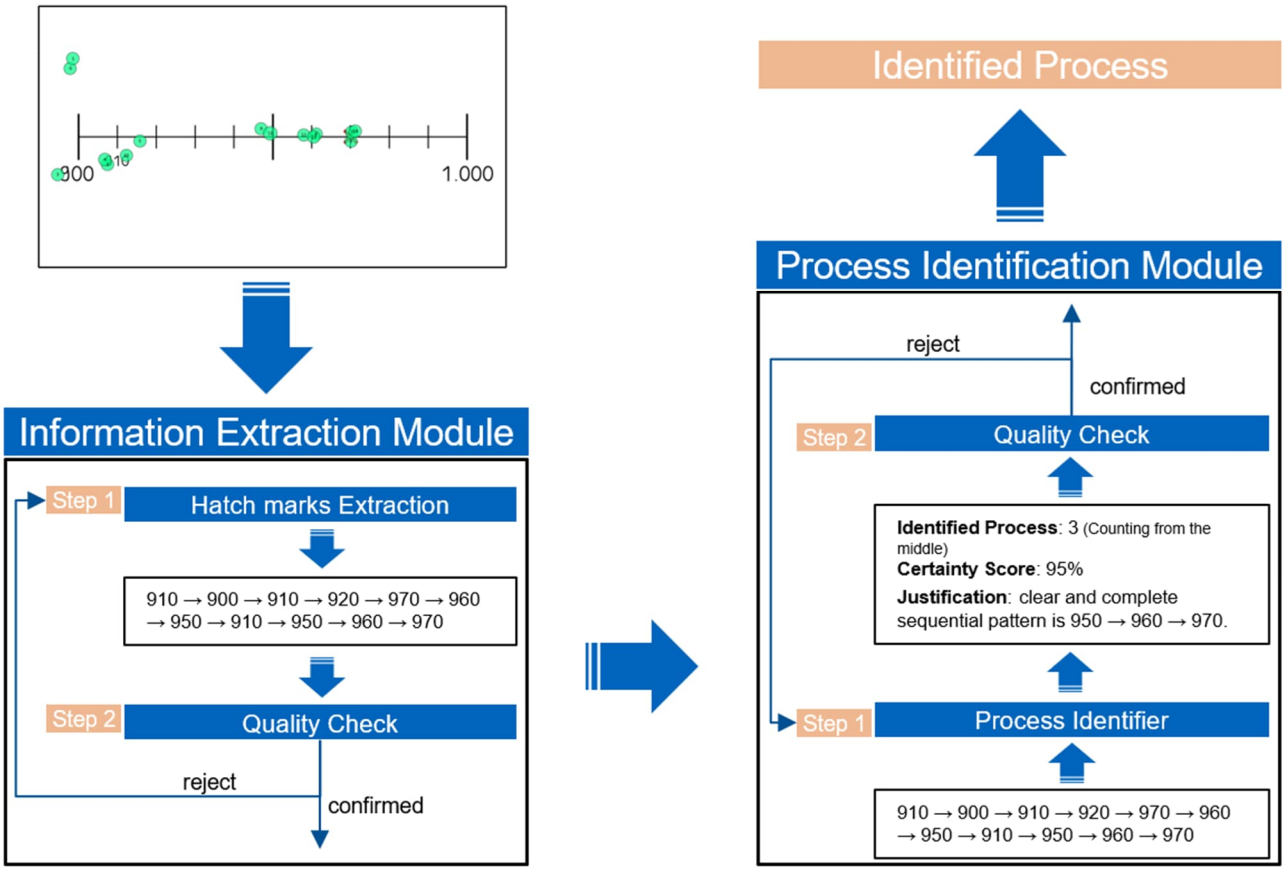
Participants and Eye tracker

This study involved 136 fifth-grade students from a German comprehensive school, who completed number line tasks. The tasks asked to locate numbers on a partially marked number line (see Fig. 2). The tasks were conducted in a controlled environment at the students' school, using the Tobii Pro X3-120 eye tracker (120 Hz, infrared, binocular) to record gaze data.

Data Collection

Fixation maps were only included when student ET accuracy was less than 2°, equivalent to approximately 2 cm on the screen. Also, only correctly solved tasks where included, since student who gave wrong answers partially just guessed or skipped that tasks, which resulted in gazes that were not meaningful for ET analyses. This way, 34 fixation maps were excluded from the analysis.

Figure 1: Pipeline for the VETA system. Input is a fixation map, and the black-bordered boxes show the VETA system input/output at each step.

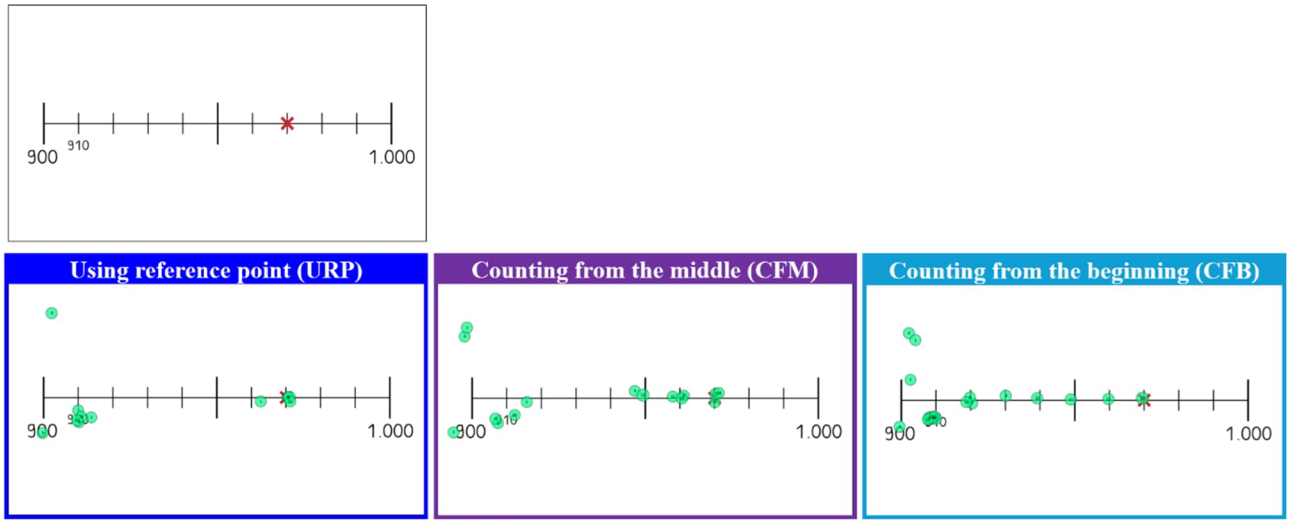


Task and Processes

In this paper, we consider number line tasks in which students are asked to name the number corresponding to a given position on the number line. The process categories used in this study are based on frameworks and findings from prior research (Simon et

al., 2023). On the number line, students can use different processes to determine which number is located at the red cross: They may rely on a reference point such as the beginning of the number line to estimate the number at the red cross (Fig. 2, left); locate the midpoint of the number line and count from there to the target (Fig. 2, middle); or count the hatch marks starting from the beginning of the number line until they reach the red cross (counting from the beginning) to find the corresponding number (Fig. 2, right).

Figure 2: The number line task and the respective processes. Using reference point (URP), counting from the middle (CFM), and counting from the beginning (CFB).



Evaluation metric and Labelling

Mathematics education researchers analysed fixation maps representations of students' ET data and assigned each gaze pattern to a predefined process. To evaluate the agreement between researchers and VETA, we used three metrics: accuracy, F1-score, and Cohen's kappa. Accuracy provides a straightforward measure of the proportion of correctly classified instances, but it can be misleading in imbalanced datasets. To mitigate this, we also report the F1-score, which balances precision and recall. Cohen's kappa measures agreement between predicted labels (VETA) and true labels (assigned by the mathematics education experts) while accounting for chance agreement, making it useful for imbalanced classes. These three metrics, offer a comprehensive picture of the model's performance.

RESULTS

In this section, we evaluate the performance of VETA pipeline on two number line estimation tasks (970 and 990), focusing on accuracy, F1-score, and Cohen's Kappa. We tested the system in two modes: In Mode 1, VETA provides a label when it detects a process with sufficient certainty; in Mode 2, it had to decide even if its confidence was low. In Mode 1 (Table 1), the system achieved perfect agreement with the expert.

Table 1: Performance of VETA in Mode 1 (labels only with sufficient certainty).

Task	Accuracy	F1-Score	Cohen's Kappa
Task 970	100%	100%	1
Task 990	100%	100%	1

A limitation of this mode is that some cases are not labelled. This may be unsuitable. For example, for task 970 the system did not provide labels for 13 out of 55 samples (~24%), and for task 990 it did not provide labels for 2 out of 47 samples (~4%).

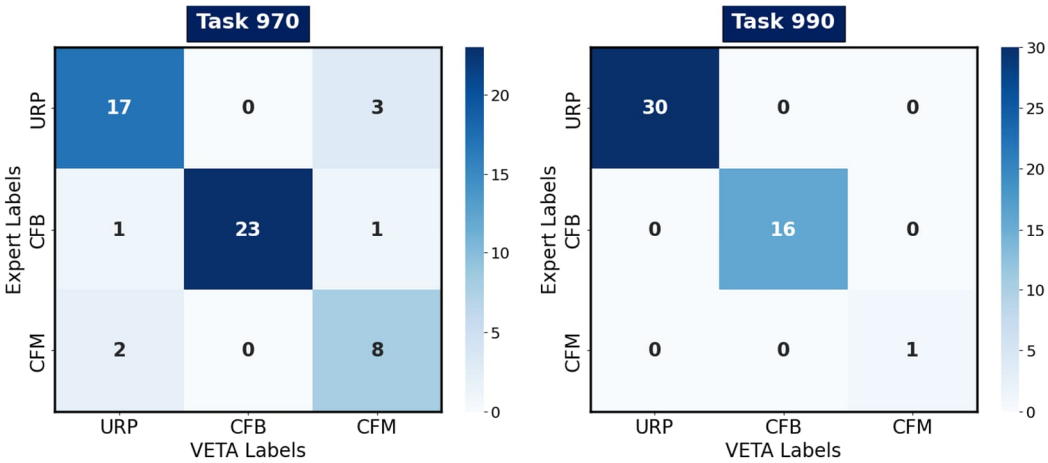
In Mode 2, the system outputs the process that best matches the sequence pattern (Table 2), even when it is not confident in the identification. For task 990, this does not affect performance. For task 970, however, we observe a reduction in performance, with a kappa of 0.82, which nonetheless reflects strong agreement with the researchers' labels.

Table 2: The performance of VETA in Mode 1 (labels for every sequence).

Task	Accuracy	F1-Score	Cohen's Kappa
Task 970	87%	86%	0.82
Task 990	100%	100%	1

To provide a more detailed analysis, we examined the confusion matrix (see Fig. 3). The x-axis represents the labels predicted by VETA, and the y-axis shows the expert-provided labels. For task 990, the performance was perfect, despite the class imbalance in the data. It is also worth noting that conventional machine learning methods cannot reliably learn from a class represented by a single sample (e.g., the "counting from the middle" process), whereas our system can handle this by integrating expert knowledge into the model. For task 970, some samples are misclassified. These errors may correspond to cases in which multiple processes are present, and no single process is clearly dominant, or to highly noisy sequences that are inherently difficult to classify.

Figure 3: Confusion matrix for the two number line tasks.



Regarding explainability, VETA provides a justification for each identified process. It explains, based on which observations in the sequence, the process was inferred. For example, as shown in Figure 1, for a specific sample, the system outputs the explanation: "The student clearly identifies the midpoint 950 and then counts forward using the sequence 950 → 960 → 970." The explanations can help educators understand students' process and see how each decision is grounded in the actual eye-movement pattern, rather than receiving only a label from the system. This also makes it easier to evaluate and validate the system's performance.

DISCUSSION

The aim of this study was to investigate how effectively a VLM-based model can analyze students' processes considering the example of a number line task using ET data. We recorded and analyzed ET data from 136 fifth-grade students and applied our VLM-based pipeline (VETA) to identify their processes based on their fixation maps.

Our results show perfect agreement for task 990 (Cohen's $\kappa = 1.00$) and strong agreement for task 970 (Cohen's $\kappa = 0.82$) compared to manual labelling by researchers. This level of accuracy suggests that an AI-system can help overcome key constraints of human analysis, such as a high time demand and the need for specialized domain expertise. It therefore opens up possibilities for integrating such analyses into digital educational applications, for example, adaptive learning systems. These findings represent an important step toward reliable and explainable adaptive learning environments, in which the automatic and accurate identification of students' processes is a prerequisite for providing meaningful individualized feedback. In addition, the system's explainable outputs can support educators in monitoring students' progress and understanding how individual's processes develop over time.

A key advantage of using VETA is that it does not require labelled data samples for training, thereby avoiding common issues in traditional machine learning models, such as imbalanced datasets and insufficient samples per class (Boels et al., 2024). Furthermore, the system explicitly leverages the temporal information in the ET data, which supports more precise process identification than approaches that rely solely on spatial gaze features (e.g., heatmap). Moreover, the VETA system is not limited to number-line tasks; it can also be used to analyze other mathematics education tasks, including quantity recognition, for example.

Despite their enormous potential for using ET in mathematics education broadly, VLM-based systems have some limitations. First, their design requires domain knowledge, and they need to be thoroughly tested to ensure robust and reliable performance. Second, their inference time may be a challenge for applications that require near real-time feedback. Third, fully validating the system remains difficult, as the inner mechanism of VLMs resemble a black box.

References

Andrá, C., Lindström, P., Arzarello, F., Holmqvist, K., Robutti, O., & Sabena, C. (2015). Reading mathematics representations: An eye tracking study. *International Journal of Science and Mathematics Education, 13,* 237–259. https://doi.org/10.1007/s10763-013-9484-y

Boels, L., Moreno-Esteva, E.G., Bakker, A., & Drijvers, P. (2024). Automated gaze-based identification of students' strategies in histogram tasks through an interpretable mathematical model and a machine learning algorithm. *International Journal of Artificial Intelligence in Education, 34*, 931–973. https://doi.org/10.1007/s40593-023-00368-9

Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., ... & Chandra, V. (2024). An introduction to vision-language modeling. *arXiv preprint arXiv*:2405.17247.

Duchowski, A. (2017). *Eye tracking methodology: Theory and practice*. Springer. https://doi.org/10.1007/978-3-319-57883-5

Hao, S., Pan, H., & Zhang, D. A. (2025). Process-oriented approach to assessing high school students' mathematical problem-solving competence: Insights from multidimensional eye-tracking analysis. *Education Sciences*, 15(6), 761. https://doi.org/10.3390/educsci15060761

Kennel, K., Ishimaru, S., Küchemann, S., Steinert, S., Kuhn, J., & Ruzika, S. (2025). Gaze-Based Prediction of Students' Math Difficulties-A Time Dynamic Machine Learning Approach to Enable Early Individual Assistance. *International Journal of Artificial Intelligence in Education*, 1-25. https://doi.org/10.1007/s40593-024-00447-5

Klein, C., & Ettinger, U. (2019). *Eye movement research: An introduction to its scientific foundations and applications.* Springer. https://doi.org/10.1007/978-3-030-20085-5

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

Simon, A. L., Asghari, P., Lilienthal, A. J., & Schindler, M. (2023). Strategy use in number line tasks of students with and without mathematical difficulties: A study using eye tracking and AI. In M. Ayalon, B. Koichu, R. Leikin, L. Rubel., & M. Tabach (Eds.), *Proceedings of the 46th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 211-218). PME.

Schindler, M., Schaffernicht, E., & Lilienthal, A. J. (2019). Differences in Quantity Recognition Between Students with and without Mathematical Difficulties Analyzed Through Eye: Analysis Through Eye-Tracking and AI. *In 43rd Conference of the International Group for the Psychology of Mathematics Education, Pretoria, South Africa*, 7-12 July, 2019 (Vol. 3, pp. 281-288). PME.

Schindler, M., Simon, A. L., Baumanns, L., & Lilienthal, A. J. (2025). Eye-tracking research in mathematics and statistics education: Recent developments and future trends. A systematic literature review. *ZDM Mathematics Education, 57,* 727–743. https://doi.org/10.1007/s11858-025-01699-8