Toward Vision-Language Models as AI Tutors in Physical STEM Manipulation

Changsoo Jung
Computer Science
Colorado State University
Changsoo.Jung@colostate.edu

Nathaniel Blanchard
Computer Science
Colorado State University
Nathaniel.Blanchard@colostate.edu

ABSTRACT

Collaborative physical manipulation tasks are widely used in STEM education to develop spatial reasoning and collaborative problem-solving. AI tutoring systems for these settings remain limited because they cannot reliably perceive student actions or evaluate in 3D whether those actions are spatially correct. This study investigates how Vision-Language Models (VLMs) can serve as perception and judgment modules in AI tutoring systems for physical block construction. Preliminary work has revealed a fundamental dissociation in current VLMs: they can identify what a student did but cannot evaluate whether it is spatially correct. Our proposed work extends this foundation toward a full AI tutoring pipeline that integrates multimodal signals and is validated through a learning outcomes study.

Keywords

vision-language models, AI tutoring, spatial reasoning, STEM education, collaborative learning

1. INTRODUCTION

This dissertation investigates the use of Vision-Language Models (VLMs) as perception and judgment modules in AI tutoring systems for collaborative physical STEM tasks. The central premise is that effective tutoring in such tasks requires automated systems that can both interpret a learner's actions on shared physical artifacts and evaluate whether those actions move the task toward its goal. Preliminary work indicates that current VLMs can identify what a student did but struggle to evaluate whether it is spatially correct, exposing a fundamental capability gap between perception and judgment. The remainder of this paper details the open problem and research challenges, the theoretical framing of the work, preliminary findings, the proposed research plan, and the expected contributions of the dissertation.

2. PROBLEM STATEMENT AND RESEARCH CHALLENGES

Changsoo Jung, and Nathaniel Blanchard. Toward Vision-Language Models as AI Tutors in Physical STEM Manipulation. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 889–891. International Educational Data Mining Society (2026).

https://doi.org/10.5281/zenodo.21039960

Collaborative physical manipulation is a cornerstone of STEM education because it develops spatial reasoning and procedural problem-solving [4, 10]. When students build a structure together, a tutor must monitor what is being built, detect deviations from the goal, and deliver corrective feedback in real time. Existing intelligent tutoring systems have focused on language-based feedback [7, 9] and symbolic domain reasoning [11]. These approaches do not transfer to physical manipulation where collaborative group dynamics directly shape how students perform actions on shared artifacts. Recent VLMs are a natural candidate for this role given their ability to process images and language jointly, but several challenges remain before VLMs can be reliably applied as AI tutors in classroom environments.

Visual complexity. Real manipulation sessions involve hands occluding blocks, inconsistent lighting, and camera perspectives that compress depth. Unlike clean rendered scenes used in prior spatial reasoning benchmarks [5, 6], a VLM processing a real video frame must reason about a noisy and partially observable 3D scene.

Incremental state change. A construction task unfolds one block at a time over 20 to 40 minutes. An AI tutor must detect a single block addition, removal, or repositioning against a growing background structure. Detecting absence (a block that was removed) is especially difficult because it requires comparing two states at a location that now appears empty.

Spatial correctness judgment. Identifying what a student did is not the same as evaluating whether it is correct. A tutor must compare the student's result against a goal configuration and determine whether the spatial relationship between blocks matches the target. This requires a form of 3D reasoning that goes beyond object recognition.

Multimodal collaboration context. In multi-party manipulation sessions, directors give spoken instructions to guide the builder [12]. An AI tutor that can ground these instructions spatially could use them to better interpret ambiguous student actions. It is unclear how much this conversational context helps VLM judgment.

These challenges together define the open problem that our study addresses: *can VLMs be made reliable enough to serve as the perception and judgment backbone of an AI tutoring system in a real physical manipulation task, and if not, what is missing?*

3. THEORETICAL FRAMING

In this study, we explore two complementary frameworks. The *model-tracing* paradigm from intelligent tutoring systems [11] holds that an effective tutor must maintain a model of the student's current state and compare it against a target. In a physical manipulation task, this means tracking a 3D block configuration and detecting deviations from a goal structure at each step. We also explore on *grounded language understanding* [2] which requires that language and vision be jointly anchored in physical space. Together, these frameworks motivate our central research question: can VLMs acquire sufficient 3D spatial grounding to support accurate tutoring feedback in physical STEM lab experiments?

4. PRELIMINARY STUDY

As the first phase of the study, we designed and ran an experiment to evaluate whether current VLMs can judge student manipulations in a physical construction task. As illustrated in Figure 1, each case provides four images: a block reference sheet, an initial state, a builder result, and a goal state. Models must identify what manipulation occurred and judge whether the builder's result matches the goal. We evaluated four frontier VLMs on 421 cases drawn from 19 real sessions [12] spanning place, move, and delete actions.

The central finding is a fundamental dissociation. Open-weight 7-8B models correctly identify what manipulation a student performed in up to 93.8% of cases when given structured state contexts. However, their judgment of whether it achieves the spatial goal remains at chance even with the same context. Proprietary frontier models (GPT-5.2 [8], Claude Sonnet 4.6 [1], Gemini 3 Flash [3]) achieve 80-93% judgment accuracy. This result shows that *knowing what happened* and *evaluating whether it is correct* are distinct capabilities in current VLMs and that the second does not follow from the first. This is the core problem we aim to resolve in the proposed studies.

5. PROPOSED RESEARCH

Study 2: Fine-Grained Spatial Error Analysis. The study 1 (Section 4) measures correctness judgment as a binary outcome. Our next step is to decompose errors into fine-grained spatial error types including position offset, stacking level error, and shape confusion. This will identify precisely where VLM spatial reasoning breaks down and inform what targeted representational support is needed for each error type.

Study 3: Multimodal Context Integration. The underlying dataset [12] contains synchronized speech and gesture annotations alongside 3D structure states. Director instructions such as "put the red block to the left of the blue one" provide explicit spatial context that may help a VLM resolve ambiguous context. We will investigate whether integrating transcribed speech and gesture improves judgment accuracy and at what cost in system complexity. This experiment extends our study to multimodal learning analytics.

Study 4: AI Tutor Prototype and Learning Outcomes. The final phase will integrate findings from Studies 1-3 into an AI tutor prototype and evaluate it in a controlled user study. Participants will complete a block construction task with or without AI tutor correctness feedback. The primary outcome is task accuracy, and the secondary measure is spatial reasoning transfer assessed through a post-task evaluation. This study will provide the ecological validity and learning science evidence.

6. EXPECTED CONTRIBUTIONS

Our current study and future research will produce four contributions to the AI and education communities. First, a benchmark for evaluating VLMs as AI tutors in real collaborative physical manipulations with reconstructable 3D ground-truth annotations. Second, empirical evidence that action identification and spatial correctness judgment are distinct VLM capabilities with clear implications for tutoring system design. Third, design guidelines for representing 3D spatial information to VLMs covering input modality and multimodal context integration. Fourth, a controlled study connecting VLM-based tutoring feedback to student learning outcomes in a physical construction task. Taken together, these contributions bridge the gap between VLM capability evaluation and practical AI tutoring design in physical STEM learning environments.

7. CONCLUSION

Physical STEM manipulation is a rich and underexplored setting for AI tutoring research at the intersection of computer vision and learning sciences. This study addresses a concrete open problem: current VLMs can recognize student actions but cannot reliably evaluate whether those actions are spatially correct. Bridging this gap requires better spatial reasoning models and a clearer understanding of what representational support AI tutors need in physical environments. Addressing this gap will advance both the theoretical understanding of VLM spatial reasoning and the practical development of AI tutoring systems for education.

8. REFERENCES

[1] Anthropic. System card: Claude Sonnet 4.6. Technical report, Anthropic, February 2026. Accessed: 2026-03-17.

[2] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.

[3] G. DeepMind. Gemini 3 flash technical report. Technical report, Google DeepMind, 2025.

[4] I. Khebour, R. Brutti, I. Dey, R. Dickler, K. Sikes, K. Lai, M. Bradford, B. Cates, P. Hansen, C. Jung, et al. When text and speech are not enough: A multimodal dataset of collaboration in a situated task. *Journal of Open Humanities Data*, 10:1–14, October 2024.

[5] F. Liu, G. Emerson, and N. Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

[6] W. Ma, H. Chen, G. Zhang, Y.-C. Chou, J. Chen, C. de Melo, and A. Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International*

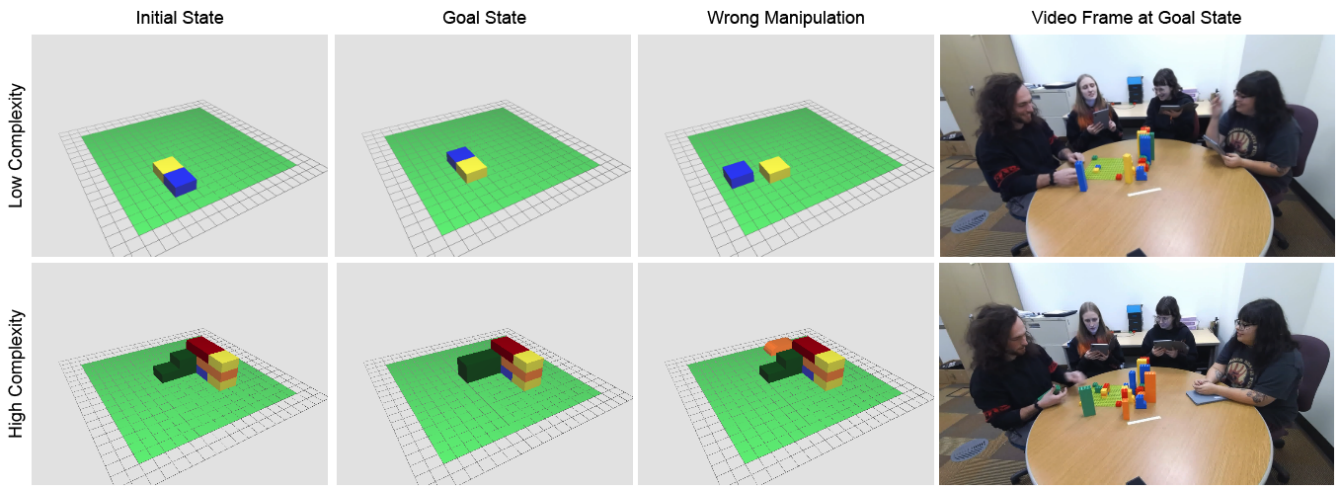

Figure 1: Examples used in the preliminary study to illustrate manipulation judgment across two complexity levels. The top row shows a low complexity structure, and the bottom row shows a high complexity structure. The four columns show the Initial State, Goal State, Wrong Manipulation, and Video Frame at Goal State. In the low complexity example, the required action is placing a square blue block next to the yellow block, but the blue block is placed in the wrong location. In the high complexity example, the required action is to place the square green block at level 2, but an orange double-sided curve block is placed at the back of the structure.

Conference on Computer Vision, pages 6924–6934, 2025.

[7] J. Macina, N. Daheim, S. Chowdhury, T. Sinha, M. Kapur, I. Gurevych, and M. Sachan. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore, December 2023. Association for Computational Linguistics.

[8] A. Singh, A. Fry, A. Perelman, A. Tart, A. Ganesh, A. El-Kishky, A. McLaughlin, A. Low, A. Ostrow, A. Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[9] A. Tack and C. Piech. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues. *arXiv preprint arXiv:2205.07540*, 2022.

[10] H. VanderHoeven, M. Bradford, C. Jung, I. Khebour, K. Lai, J. Pustejovsky, N. Krishnaswamy, and N. Blanchard. Multimodal design for interactive collaborative problem-solving support. In *International Conference on Human-Computer Interaction*, pages 60–80. Springer, 2024.

[11] K. VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011.

[12] Y. Zhu, C. Jung, K. Lai, V. Venkatesha, M. Bradford, J. Fitzgerald, H. Jamil, C. Graff, S. K. G. Kumar, B. Draper, N. Blanchard, J. Pustejovsky, and N. Krishnaswamy. Multimodal common ground annotation for partial information collaborative problem solving. In *Proceedings of the 21st Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-21)*, pages 85–91, Düsseldorf, Germany, Sept. 2025. Association for Computational Linguistics.