Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Designing an AI-Enhanced Timeline for Monitoring Multimodal Interactions in Embodied Learning Environments

Joyce FONTELES[a*], Namrata SRIVASTAVA[ab], Eduardo DAVALOS[a], Ashwin T S[a] & Gautam BISWAS[a]
[a]*Vanderbilt University, Nashville, United States*
[b]*Center for Learning Analytics, Monash University, Australia*
*joyce.h.fonteles@vanderbilt.edu

Abstract: Embodied learning represents a natural and immersive approach to education, where the physical engagement of learners plays a critical role in how they perceive and internalize concepts. This method allows students to actively embody and explore knowledge through interaction with their environment, significantly enhancing retention and understanding of complex subjects. However, researchers face significant challenges in exploring children's learning in these physically interactive spaces, particularly due to the complexity of tracking multiple students' movements and dynamic interactions in real-time. To address these challenges, this paper introduces a Double Diamond design thinking process for developing an AI-enhanced timeline aimed at assisting researchers in visualizing and analyzing multimodal interactions within embodied learning environments. We outline key considerations, challenges, and lessons learned in this user-centered design process. Our goal is to create a timeline that employs state-of-the-art AI techniques to help researchers interpret complex datasets, such as children's movements, gaze directions, and affective states during learning activities, thereby simplifying their tasks and augmenting the process of interaction analysis.

Keywords: Embodied learning, Multimodal Learning Analytics, Artificial Intelligence

1. Introduction and Motivation

Embodied learning represents a natural and immersive approach to education, where the physical engagement of learners plays a critical role in how they perceive and internalize concepts. By actively involving their bodies, learners can interact with their environment in ways that significantly enhance retention and understanding of complex subjects (Danish et al., 2020). For instance, Lindgren et al. (2016) demonstrated that students who physically acted out the movement of meteors in a game had a deeper understanding of object motion and the influence of gravity compared to those who used a traditional desktop simulation. They found that when learners enact concepts and experience critical ideas through whole-body interactions, it not only leads to significant learning gains, but also builds a more positive attitude towards science.

Monitoring and measuring embodied learning are essential to better understand how students interact with these physical learning environments. In contrast to conventional learning setups where verbal or written interactions, or clickstreams, are easy to track, embodied learning involves capturing a range of complex non-verbal cues, including body movements, spatial positioning, gaze direction, and emotional expressions. These elements are crucial for understanding how learners engage and collaborate in real-time.

Traditional methods for measuring embodied learning have relied on qualitative approaches like Interaction Analysis (IA), where human observers manually analyze classroom footage and interactions to identify patterns in learning behavior. While effective, these methods are highly resource-intensive, requiring significant time and human effort to

process data (Zhou et al., 2024). Advances in artificial intelligence and multimodal learning analytics enable the automated tracking of students' movements, gaze, and emotional states, offering a more scalable and efficient solution. AI-based systems can interpret complex datasets and provide insights that are otherwise difficult to capture in real-time through traditional observation alone. Despite these advances, current challenges remain in scaling these methods to real-world classrooms. This paper proposes an AI-enhanced timeline designed to assist researchers in visualizing and analyzing multimodal interactions within embodied learning environments. By leveraging AI, this system aims to simplify the process of monitoring and interpreting students' activities within embodied learning contexts.

Previous works on visualizations of multimodal data, such as Ez-Zaouia's Emodash (Ez-zaouia et al., 2020), have explored the visualization of learner emotions and system interactions in online learning sessions. Emodash, for example, reinforced the challenge of identifying the right level of detail and timescales for effective visualizations. This work builds upon previous research that explored visualizing learners' performances and behaviors using primarily systems logs dashboards, as well as the design of multimodal and contextual emotional dashboards for tutors (Schwendimann et al., 2017). Similarly, Fernandez-Nieto et al. (2021) developed interactive timelines to integrate multimodal data, such as physiological responses and logged actions, in clinical simulations, leveraging data storytelling principles to aid students' reflection during debriefing sessions.

Our work extends these into the domain of embodied learning, where real-time interactions in a 3D space add layers of complexity to data capture and analysis. We aim to augment the capabilities of learning scientists in making inferences by providing a more dynamic, real-time view of student interactions. Our timeline not only integrates system interactions and affective data but also incorporates gaze tracking to study attention shifts during learning activities. This is particularly crucial in embodied learning environments, where physical movement, gaze, and emotional engagement are integral to the learning process. Through synchronizing the timeline's multimodal data with video playback, we allow researchers to focus on high-level interpretation while AI handles lower-level data inferences. This approach empowers human researchers to interpret the learning process while still benefiting from automated assistance, making it a valuable tool for IA in complex embodied environments.

2. Timeline System Design

2.1 Design process

The development of the timeline followed a "human-in-the-loop" co-design process. We actively engaged end-users, ensuring a profound comprehension of their needs and the tasks they aim to achieve with the timeline. Specifically, our collaboratively multidisciplinary research team consists of experts with technical skills (computer scientists (CS)) and primary users of the timeline (learning scientists (LS)). The learning scientists have experience with the Interaction Analysis methodology for assessing embodied learning data through qualitative observations.

We initiated the development process by following the Double Diamond (DD) design framework. The DD framework[1] consists of two main phases—divergent and convergent thinking (Van Tyne, 2022)—which occur twice throughout the process, encompassing four key stages: Discover, Define, Develop, and Deliver (as shown in Figure 1).

In the **Discover** phase, we took diverging approaches to gain a comprehensive understanding of the challenges involved in collecting and analyzing multimodal data in embodied learning environments. Specifically, we employed two diverging approaches: a controlled lab pilot study with graduate students and an analysis of archival data from real classroom settings, where children participated in embodied learning activities. By combining these two complementary approaches, we were able to develop a more thorough

[1] https://www.designcouncil.org.uk/our-resources/framework-for-innovation

understanding of the data and its context across diverse settings.

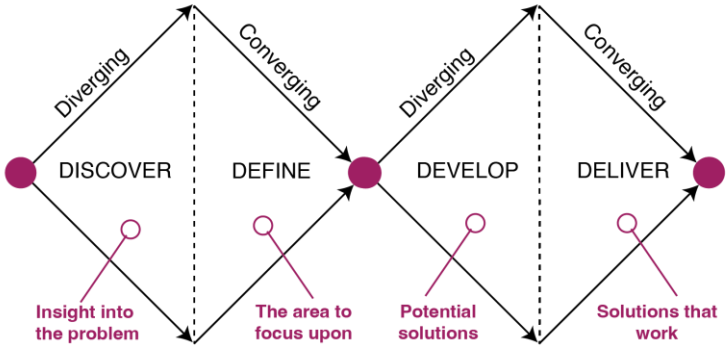

Figure 1. The Double Diamond Design Thinking Process

Next, in the **Define** phase, we held discussions with learning scientists to examine their current methodologies for analyzing multimodal data in embodied learning environments. This process involved reviewing the tools and techniques they most commonly use to capture and interpret student interactions. Our goal was not only to understand how these methods support their analysis but also to identify where they fall short in capturing the full complexity of embodied learning activities. By assessing both the strengths and limitations of these approaches, we aimed to pinpoint areas where AI-driven methods could enhance the analysis process and help researchers manage the large volumes of multimodal data more efficiently.

Following the completion of the first design cycle, the second cycle begins with the Develop phase. In this phase, we explored various data analysis and visualization techniques to accurately represent the multimodal data generated in embodied learning environments. This exploration was a collaborative effort between LS and CS team, with each team contributing their expertise and offering unique perspectives to guide the development process.

Finally, in the **Deliver** phase, we evaluated the solutions, refined them, and ensured that the final design met the evolving requirements of the LS team. This phase involved iterative testing with end-users, such as learning scientists, to gather feedback on the usability and effectiveness of the AI-enhanced timeline.

2.2 System Architecture

2.2.1 Synchronous MMLA data collection

Synchronizing multiple data streams during classroom activities is a critical challenge in analyzing multimodal interactions in real-time learning environments. In our study, it was essential that the data—such as video from different camera angles, audio, system logs, and simulation recordings—were collected in a time-aligned manner to ensure accurate interaction analysis. Time-aligned data is crucial for creating a timeline visualization where videos from multiple angles can be analyzed with added context, and inferences can be derived from the multimodal data without needing to manually synchronize the streams, which is time-consuming and prone to error.

To address this challenge, we employed ChimeraPy (Davalos et al., 2023), an open-source distributed streaming framework. It enabled us to rapidly deploy in the classroom, ensuring that multiple sources of data were synchronized from the moment of collection. This eliminated the need for manual post-collection alignment, which is both resource-intensive and a potential source of inaccuracies.

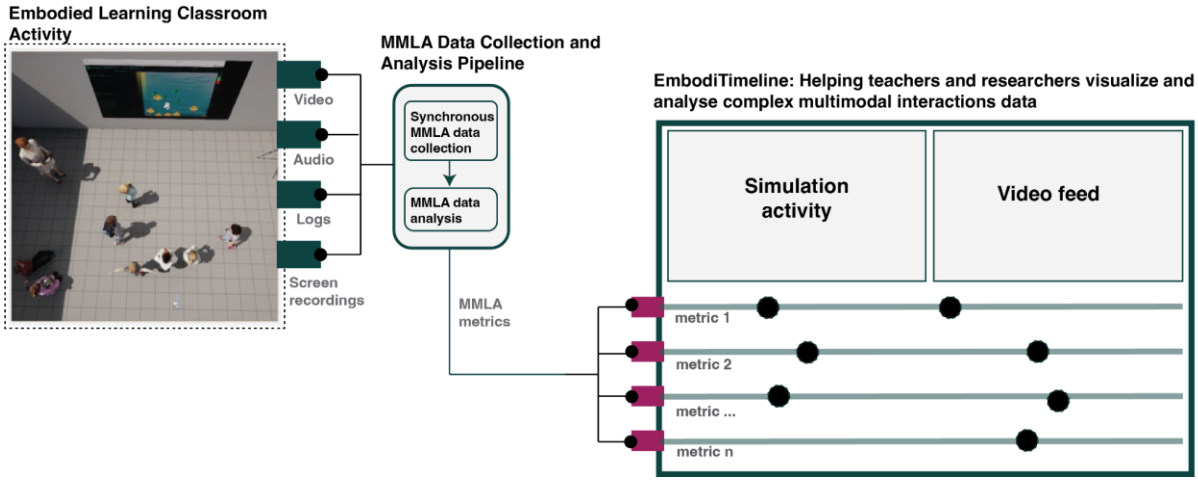

Figure 2. Architecture shows the processes of data collection in classroom, MMLA analysis and Timeline Visualization

Thus, our timeline design builds upon this pre-synchronized data to support Interaction Analysis effectively. IA typically requires researchers to spend significant amounts of time logging content and manually aligning data from multiple sources. The timeline is fully interactive, allowing researchers to click on any segment or row to jump directly to specific points in the video, providing targeted insights based on the multimodal data being displayed, while displaying an integrated view of student behavior during the embodied learning activity.

2.2.2 MMLA data analysis

The analysis of system logs was essential for understanding the nuances of student interactions and their progression within the MR environment. These logs provided a temporal record of molecule changes and movements made by students throughout the learning activity, allowing researchers to identify patterns in how students navigated the virtual space, interacted with different elements, and transitioned between molecules. The dataset yielded three key dimensions: (1) Students' States—tracking which molecules students embodied at different times helped explore their evolving understanding of scientific concepts; (2) Students' Actions—analyzing actions associated with molecular embodiment was fundamental to gauging students' grasp of the photosynthesis process; and (3) System State—capturing whether the simulation was in daylight or night.

Figure 3 illustrates the pipeline for processing video frames from multiple camera angles, allowing us to perform person reidentification, affect detection, and gaze estimation. The process begins with video frames from multiple cameras, which are passed through an MTCNN to detect faces. Using the HSEmotions model for facial landmark detection enabled us to capture detailed facial expressions and predict valence and arousal scores. These two steps result in face crops, bounding box coordinates, and valence-arousal scores for each frame of every camera. These outputs feed into our person reidentification (ReID) algorithm, which links faces across frames and cameras, resolving issues like occlusion or students leaving the camera's view. This results in face tracklets, where each bounding box and face crop is attached to a specific person ID for every frame.

Affect detection then leverages the person tracklets and the valence-arousal scores. These scores were categorized based on Russell's circumplex model of emotions (Russell, 1980) and D'Mello's dynamics of affective states (D'Mello & Graesser, 2012). The emotions are categorized into both learning-centered emotions, which are specific to educational contexts, and quadrant-based emotions based on the four quadrants of Russell's model.

Using the face tracklets from the first line, L2CSNet estimates gaze vectors based on the monocular images from each camera. However, these vectors are in 2D camera coordinates. To understand where students are looking in the 3D environment, we apply the

ZoeDepth model to reconstruct the scene in 3D. The 2D gaze vectors are then reprojected into this 3D scene, allowing us to determine where the students are directing their gaze in that coordinate system. Having done annotations of objects of interest (OOIs) via Vision6D[2], ray tracing is then used to detect which OOIs (e.g., the screen, teacher, researcher, or other students) the gaze vectors intersect with, providing insight into students' attention within the learning environment. For more details about the technical components, please refer to (Fonteles et al., 2024).

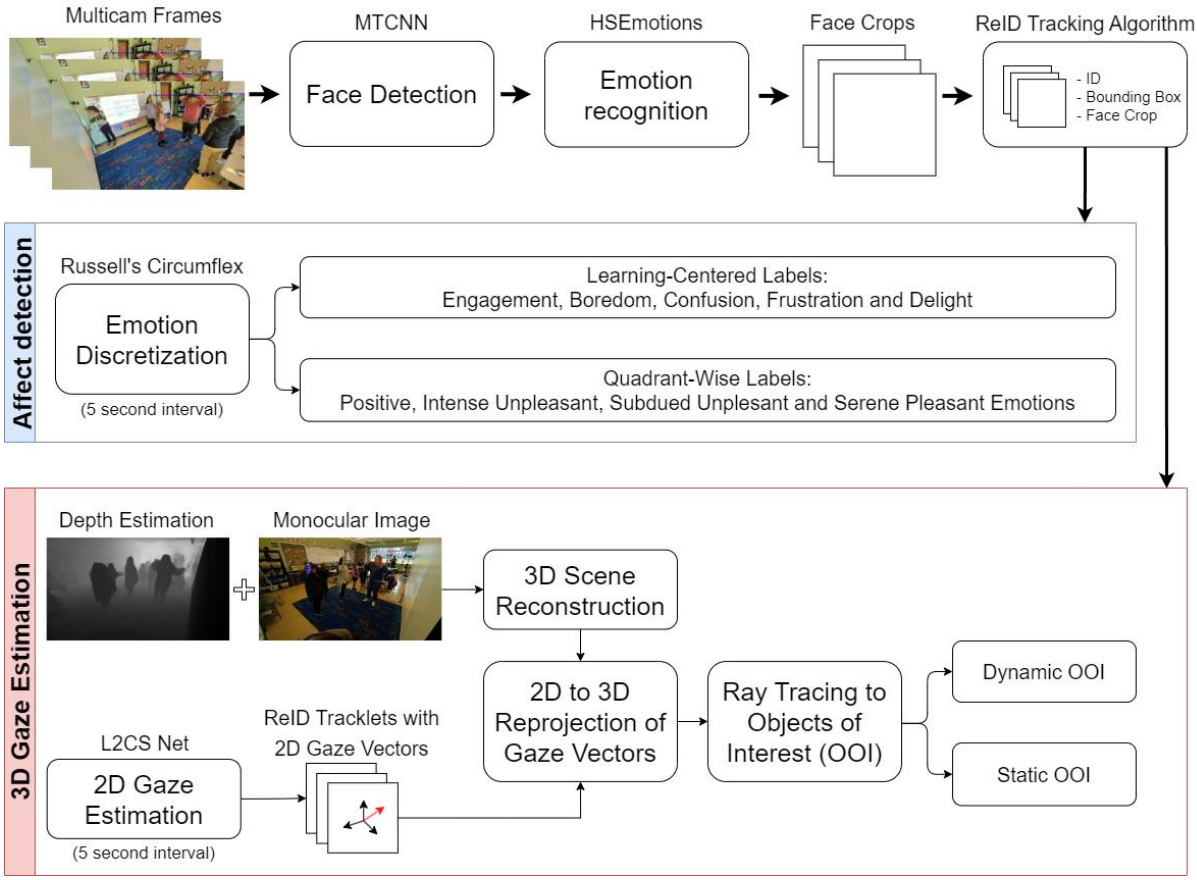

Figure 3. Pipeline for person reidentification, affect detection, and 3D gaze estimation

3. Study Context

This study is part of the GEM-STEP project (Generalized Embodied Modeling to support Science through Technology Enhanced Play), where motion-tracking technologies and mixed-reality environments enable students to embody scientific phenomena. In our study, 4th-grade students from a public school in the southeastern United States participated in a two-month project focusing on food webs and photosynthesis. We focused on the photosynthesis model to guide the tool's initial development. This closed-loop simulation allows students to repeatedly test molecular transformations as the screen alternates between day and night. The simulation features a tomato plant, a mouse, and zoomed-in views of chloroplasts and roots. As students move around the classroom, motion-tracking technology reflects their movements on the screen, allowing them to embody molecules (oxygen, carbon dioxide, water, and sugar) and interact with features in the simulation, as shown in Figure 4. The goal is for students to understand that sunlight enables carbon dioxide and water to turn into oxygen and sugar in the plant's chloroplast, driving photosynthesis. This process fuels plant growth

[2] https://github.com/InteractiveGL/vision6D

and illustrates how animals, like the mouse, depend on oxygen produced by plants, emphasizing collaboration and matter rearrangement in sustaining life.

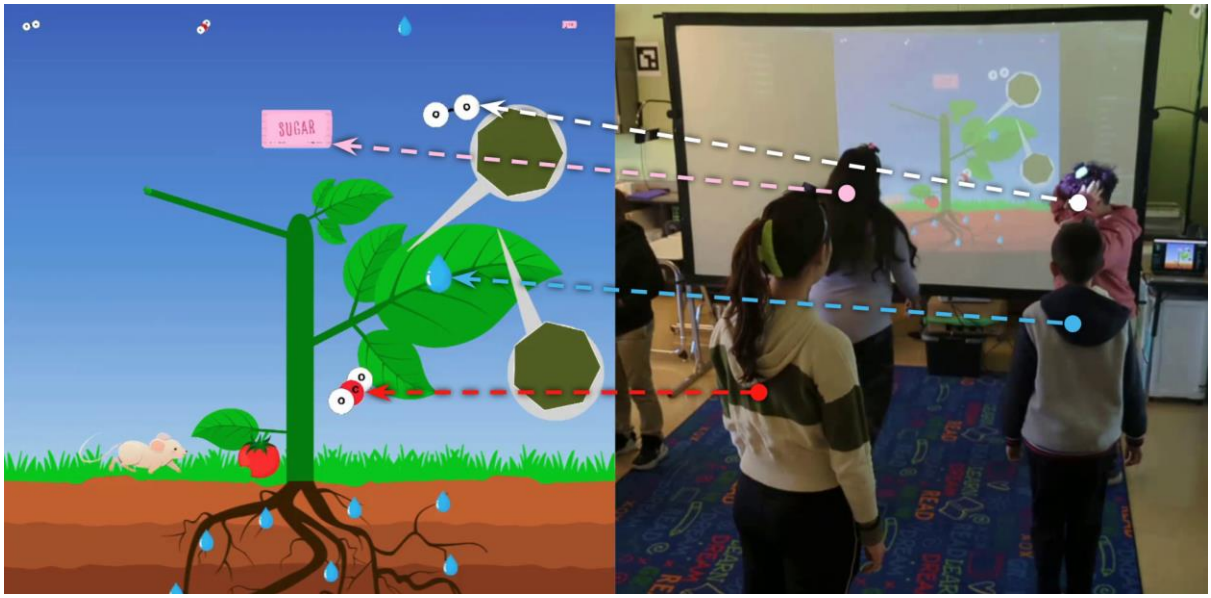

Figure 4. The photosynthesis model where children embody molecules

4. Results from design thinking process

The first diamond thinking process: Discover and Define Phase

We started the discovery process by conducting a small pilot study with 9 graduate students in a lab-based environment. The learning science researchers helped us conduct the study. The aim of this study was to build knowledge regarding the study context, understand what type of data usually collected, what metrics could be derived from the data using AI techniques, and what kinds of constraints would occur during the in-the-wild classroom studies.

Upon getting a better understanding of the dataset, we conducted regular (e.g., weekly) interaction meetings with the learning science researchers to understand what data analysis tools they utilize and what challenges are faced by them. We discovered that they predominantly use Interaction Analysis (IA) as their primary method for interpreting collaborative, embodied learning environments. IA is highly effective in revealing deep insights and capturing nuanced interactions from video data. However, the researchers noted that while IA provides valuable insights, the manual process is inherently time-consuming and requires significant human resources.

The second diamond thinking process: Develop and Deliver Phase

The learning science researchers mentioned that they would like to see a tool that can capture the nature of embodiment, i.e., provide insights on how students move across the space and how their attention shifts during contextualized learning process. Having already acquired knowledge on the types of information we could derive from embodied activities through the pilot study, we proceeded with data collection in a 4th grade science class, thus being able to assess how our models would fare with real, in-the-wild data of children's embodied experiences.

Having time-aligned data collected from system logs, screen recordings, videos and audio of the environment, we proceeded with processing this multimodal dataset to highlights students' contextualized choices within the learning environment, highlighting moments where they were able to successfully perform the correct transformations defined in the photosynthesis model, as well as all their exploratory actions beforehand since researchers were interested in investigating patterns of success and productive failures. System logs provided a temporal record of student interactions with the environment, allowing us to trace

molecule changes and movements, which revealed patterns in navigation and transitions between molecules and offered valuable context for understanding students' explorations and choices.

Timeline prototype:

Based on the insights from 4 learning science researchers during the design process, we created a basic prototype of the timeline in order to present the multimodal findings contextualized with the science model and the video of students' embodied activities. The timeline visualization, shown in Figure 5, is divided into multiple lines, each representing a different data modality captured during the photosynthesis simulation. This multimodal approach allows researchers to analyze and interpret the students' interactions with the mixed-reality environment effectively. Each segment on the timeline is fully interactive—clicking on any part of the timeline will jump the video cursor to that specific moment, providing synchronized video playback to support deeper analysis.

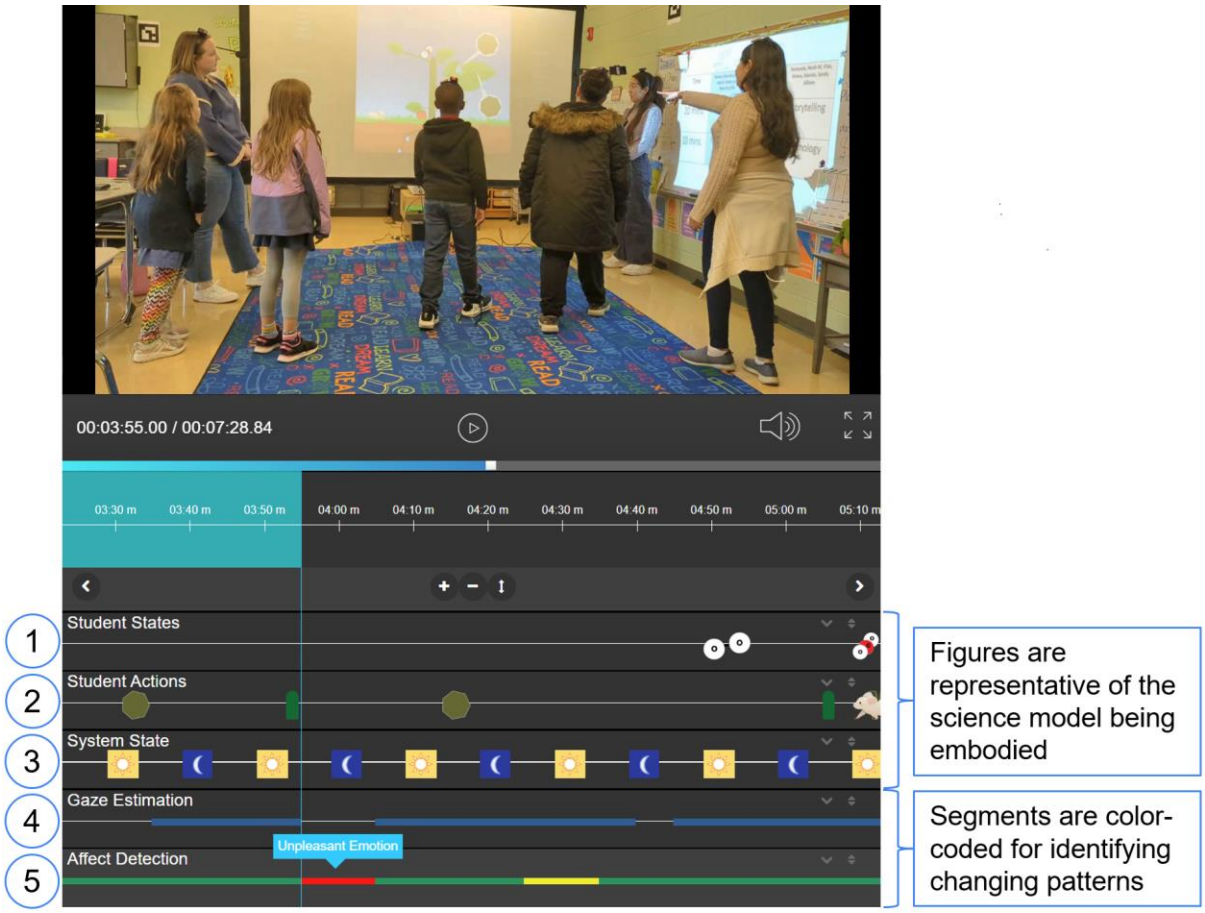

Figure 5. A basic prototype of the timeline of multimodal data

1. **Student States**: This line shows which molecule each student embodied at any given moment in the simulation (e.g., oxygen, carbon dioxide, water, or sugar). The visual representation of these molecules is consistent with what the students see in the simulation, making it easier to track which role a student played during key interactions.
2. **Student Actions**: This line represents the movements of the students within the simulation, indicating which element of the simulation they interacted with, such as the mouse, roots, the leaf's chloroplast, or the plant stem. These actions are crucial for observing molecular transformations, which only occur if the student interacts with the element that causes transformations to their current molecule.
3. **System States**: This line displays the simulation's transitions between day and night. These shifts are essential for photosynthesis, as the process occurs only during the

daytime when sunlight is present. This allows researchers to track whether students recognize that sunlight is necessary for photosynthesis.

4. **Gaze Estimation**: This line is represented by color-coded segments, showing where students directed their attention over time. It tracks whether they were looking at the screen (simulation), the teacher, another student, or the researcher. This helps researchers observe how the students' focus shifts during the activity.
5. **Affect Detection**: This line shows students' emotional states, represented by color-coded segments. The emotions are derived from facial expression analysis, which calculates valence and arousal and maps them into labeled emotions. This provides insights into how students' emotional engagement fluctuates during the activity.

For Student States, Student Actions, and System States we used figures representative of the photosynthesis model to simplify tracking and make it easier for researchers to follow the students' progress and interaction with the system. For Gaze Estimation and Affect Detection we employed color-coded segments, allowing researchers to quickly pinpoint shifts in patterns of attention and emotional engagement over time.

5. Lessons learned and future directions

Throughout the development of the multimodal timeline, we engaged in regular feedback sessions with 7 learning scientists. These sessions were integral in helping us identify challenges and opportunities for improvement, and reflect on both the usability of the system and its alignment with the researchers' needs. Based on our initial analysis during the early phases of the Double Diamond (DD) design process, it became apparent that the solutions generated were insufficient to fully address the design requirements of the timeline.

Researchers initially expressed concerns about cognitive load when presented with all multimodal findings simultaneously. They noted that the relevance of each modality might vary depending on the research question being explored, and suggested that the ability to toggle between viewing all data at once or focusing on specific subsets would be helpful. Additionally, learning scientists, who are not AI specialists, emphasized the need for AI explainability. This would help them better understand AI's role and limitations in generating insights, allowing them to more confidently interpret the findings and reflect on how human judgment and AI interact in the analysis process.

Another insight was the recognition of speech as a crucial modality for future iterations. Researchers observed that shifts in student gaze or attention were often triggered by verbal interactions—such as remarks from the teacher or peers—that influenced student actions and understanding. Including speech data would help disambiguate such shifts and provide a clearer picture of the underlying factors driving changes in behavior and learning.

Looking ahead, we plan to enhance the timeline to not only support individual student analysis but also facilitate collaborative learning analysis. We aim to provide researchers with the ability to query specific moments of interest, allowing them to dive deeper into reasoning and uncover patterns across multiple days of embodied learning—something not feasible within the constraints of traditional Interaction Analysis alone.

Acknowledgements

This work was supported by the following grants from the National Science Foundation (NSF): DRL-2112635, IIS-1908632 and IIS-1908791. The authors have no known conflicts of interest to declare. We would like to thank all of the learning scientists, students and teachers who participated in this work. We would also like to thank all the people who prepared and revised previous versions of this document.

References

Alfredo, R., Echeverria, V., Jin, Y., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). SLADE: A Method for Designing Human-Centred Learning Analytics Systems. *Proceedings of*

the 14th Learning Analytics and Knowledge Conference, 24–34. https://doi.org/10.1145/3636555.3636847

Danish, J. A., Enyedy, N., Saleh, A., & Humburg, M. (2020). Learning in embodied activity framework: A sociocultural framework for embodied cognition. *International Journal of Computer-Supported Collaborative Learning*, *15*(1), Article 1. https://doi.org/10.1007/s11412-020-09317-3

Danish, J. A., Enyedy, N., Saleh, A., Lee, C., & Andrade, A. (2015). *Science Through Technology Enhanced Play: Designing to Support Reflection Through Play and Embodiment*.

Davalos, E., Timalsina, U., Zhang, Y., Wu, J., Fonteles, J. H., & Biswas, G. (2023). ChimeraPy: A Scientific Distributed Streaming Framework for Real-time Multimodal Data Retrieval and Processing. *2023 IEEE International Conference on Big Data (BigData)*, 201–206. https://doi.org/10.1109/BigData59044.2023.10386382

Davis, B., Tu, X., Georgen, C., Danish, J. A., & Enyedy, N. (2019). The impact of different play activity designs on students' embodied learning. *Information and Learning Sciences*, *120*(9/10), 611–639. https://doi.org/10.1108/ILS-08-2019-0081

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

Enyedy, N., & Danish, J. (2014). Learning physics through play and embodied reflection in a mixed reality learning environment. In *Learning technologies and the body* (pp. 97–111). Routledge.

Ez-zaouia, M., Tabard, A., & Lavoué, E. (2020). EMODASH: A dashboard supporting retrospective awareness of emotions in online learning. International Journal of Human-Computer Studies, 139, Article 102411. https://doi.org/10.1016/j.ijhcs.2020.102411

Fernandez-Nieto, G.M., Echeverría, V., Shum, S.B., Mangaroska, K., Kitto, K., Palominos, E., Axisa, C., & Martínez-Maldonado, R. (2021). Storytelling With Learner Data: Guiding Student Reflection on Multimodal Team Data. *IEEE Transactions on Learning Technologies, 14, 695-708.* https://doi.org/10.1109/TLT.2021.3131842

Fonteles, J., Davalos, E., Ashwin, T. S., Zhang, Y., Zhou, M., Ayalon, E., Lane, A., Steinberg, S., Anton, G., Danish, J., Enyedy, N., & Biswas, G. (2024). A First Step in Using Machine Learning Methods to Enhance Interaction Analysis for Embodied Learning Environments. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial Intelligence in Education* (Vol. 14830, pp. 3–16). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64299-9_1

Lindgren, R., Tscholl, M., Wang, S., & Johnson, E. (2016). Enhancing learning and engagement through embodied interaction within a mixed reality simulation. *Computers & Education*, *95*, 174–187. https://doi.org/10.1016/j.compedu.2016.01.001

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Schwendimann, B. A., Rodríguez-Triana, M. J. , Vozniuk, A., Prieto, L. P., Boroujeni, M. S., Holzer, A., Gillet, D. & Dillenbourg, P. (2017). Perceiving Learning at a Glance: A Systematic Literature Review of Learning Dashboard Research. *IEEE Transactions on Learning Technologies, vol. 10, no. 1, 30-41.* https://doi.org/10.1109/TLT.2016.2599522

Van Tyne, S. (2022, Assessed). *Design Thinking: Divergence and Convergence Cycles*. https://www.seanvantyne.com/2017/02/19/design-thinking-divergence-convergence-cycles/

Zhou, M., Fonteles, J., Danish, J., Davalos, E., Steinberg, S., Biswas, G. & Enyedy, N. (2024). Exploring Artificial Intelligence Supported Interaction Analysis. *Proceedings of the 18th International Conference of the Learning Sciences - ICLS 2024*, 2327–2328. https://doi.org/10.22318/icls2024.926221