Koh, E., et al. (Eds.) (2026). Proceedings of the 2nd International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education.

Proactive Group Learning Design Support Through Multimodal Evidence from VR Environments

Changhao LIANG[a*], Shengkai LIN[b] & Hiroaki OGATA[c]
[a] *Research Institute for Information Technology, Kyushu University, Japan*
[b] *Graduate School of Information Sciences, Hiroshima City University, Japan*
[c] *Academic Center for Computing and Media Studies, Kyoto University, Japan*
* liang.changhao.507@m.kyushu-u.ac.jp

Abstract: Virtual reality (VR) environments generate rich multimodal evidence that informs educational design. However, existing multimodal learning analytics either rely on black-box predictive models that lack interpretability or focus on descriptive post-hoc analyses that provide insufficient support for proactive instructional design. This study aims to forge new pathways for leveraging multimodal evidence in proactive group learning design. Facial activation data from 20 undergraduate participants were collected during an immersive lecture session. Aggregated emotional intensity indicators and pairwise temporal similarity of confusion were jointly optimized through a genetic algorithm to form heterogeneous groups. Beyond grouping, pairwise temporal analysis reveals divergence and alignment patterns that inform role assignment and instructional adjustment. The illustrated case demonstrates how interpretable multimodal indicators can guide proactive, theory-informed group learning design.

Keywords: multimodal data, group learning, group formation, pairwise similarity, VR environment

1. Introduction and foundation

Immersive and blended learning settings are increasingly equipped with metaverse-based technologies. Virtual reality (VR) offers a sense of presence and interactivity that has often been missing in conventional online education (Kim et al., 2026). Beyond enhancing engagement, VR sensors also capture rich multimodal traces such as attention, stress, and emotions, opening new possibilities for adaptive learning designs.

In individual learning contexts, multimodal process data have been used to proactively estimate learning outcomes. For example, high-dimensional facial expression channels have been incorporated into deep learning models for prediction (Daza et al., 2025). Despite their effectiveness, multi-layer neural networks are often too complicated to interpret (Khosravi et al., 2022), which curbs theoretical grounding and reduces adoption in real-world practice for non-technical educators. In contrast, another line of research underscores descriptive triangulation across multiple modalities through post-hoc analyses to identify behavioral patterns and learner profiles (Tao et al., 2025). Albeit its contributions to theoretical understanding, it provides limited support for real-time instructional decision-making.

When extended to group learning, research on the unfolding of learning mechanisms with vintage theories proliferates yet remains largely descriptive triangulation of observed facts and lacks support for designing upcoming group activities. For proactive interventions such as group formation, input indicators still adopt common strategies such as aggregating multi-modalities into composite individual-level indicators, which may simplify implementation but risk compressing the richness of the original signals in immersive environments. In the CSCL community, interpretability of the underlying mechanism remains a key concern (Reimann & Baker, 2025), which makes purely black box approaches difficult to justify. VR environments provide unique modalities beyond log data that count in dissecting interpersonal interactions

(Giannakos & Cukurova, 2023), such as emotional experiences that exhibit significant intra-individual variability (Saqr, 2024), to which group learning is rather sensitive. Therefore, how to make the best of multimodal evidence for proactive data-driven group learning design while maintaining interpretability and contextual relevance remains a challenge.

This paper addresses these gaps by applying multimodal evidence to group formation. We fuse intensity and structural indicators and operationalize pairwise temporal metrics as group awareness tools to support proactive, theory-informed group learning design.

2. Multimodal evidence-based group formation and learning design

2.1 Data collection and indicator definition

The multimodal evidence in this study was collected from 20 undergraduate participants completing a VR-based learning activity consisting of lecture video watching (around 12 mins) followed by a quiz. The lecture is about educational data analysis, including an introductory segment (around 2.5 mins) and core conceptual explanations with two substantiated topics (around 3.5 and 5.5 mins respectively).

Using the SMARTe-VR platform developed with the OpenXR SDK (Daza et al., 2025), 51 facial features based on the PICO 4 face-tracking APIs were recorded at 30 Hz throughout the video-watching session with a 0-1 intensity scale. In light of Action Units (AUs) defined in the Facial Action Coding System (FACS) (Cohn et al., 2007), three AU indicators were selected for group formation. Table 1 presents the selected AUs with construct, their corresponding face indexes from the tracking APIs and grouping strategies.

Table 1. *Facial AUs for multimodal group formation and their pedagogical rationales*

AU Code (Construct)	**Face Index***	**Indicator (Distance Type for Grouping)**	**Pedagogical Rationale**
AU4 (Confusion / Cognitive difficulty)	#16 (BrowDown_L)	Aggregated (intensity-based)	Avoids uniformly low-confusion groups
		Pairwise (temporal-dynamics)	Aligns diverse confusion timing
AU45 (Boredom / Attention fluctuation)	#28 (EyeBlink_L)	Aggregated (intensity-based)	Prevents uniformly low-attention groups
AU6 (Positive engagement / Smile)	#29 (CheekSquint_L)		Ensures presence of positively engaged members to sustain interaction
AU12 (Positive engagement / Smile)	#19 (MouthSmile_L)		

** Only the left-side (L) signal was considered as the right-side data showed identical values.*

2.2 Multimodal evidence-based heterogeneous group formation

As a fundamental mechanism for proactive data-driven group learning support, group formation was conducted based on aforementioned data. To incorporate relevant constructs beyond conventional cognitive learner models, both *aggregated individual* features and *pairwise temporal* similarity were integrated into the optimization process. All selected indicators were incorporated under a heterogeneous grouping strategy that maximizes the differences among group members, in line with the pedagogical rationales in Table 1.

Within-group heterogeneity was operationalized as the fitness value (F) in a genetic algorithm-based group formation approach (Liang et al., 2025). Two distance types were defined: (1) *intensity-based* dispersion derived from aggregated activation features and (2) *temporal-dynamics* distance derived from pairwise facial activation time series.

For (1), four activation intensity indicators in Table 1 were standardized across learners using z-score normalization. Group-level heterogeneity was computed as the sum of squared deviations between individual standardized scores and the corresponding group means. For (2), facial activation (#16) time series were downsampled to 1 Hz and z-score normalized within each learner. Pairwise temporal dissimilarity was computed using band-constrained (±2

s) Dynamic Time Warping (DTW) (Ding et al., 2008). The resulting DTW distance matrix was further z-score standardized across all learner pairs. Group-level heterogeneity was then defined as the mean standardized DTW distance among all within-group pairs.

The final fitness function $F = F(aggregated) + \lambda F(pairwise)$ combined the two distance components. λ was set to 4 to balance the four aggregated indicators and doubled ($\lambda = 8$) to underscore temporal divergence for subsequent pairwise analysis. To this end, five groups of four learners were created by maximizing overall F in one trial grouping. The grouping results with aggregated feature distribution of each group are shown in Figure 1. As shown in Table 2, *F(aggregated)* are almost balanced across groups, while Group 4 has the highest overall overall (F) and *F(pairwise)* heterogeneity focus on the temporal dynamics.

Table 2. *Aggregated, pairwise, and overall fitness values denoting group heterogeneity*

	Group 1	**Group 2**	**Group 3**	**Group 4**	**Group 5**
F(aggregated)	2.354	2.609	**2.874**	2.635	2.665
F(pairwise)	0.492	0.416	0.212	**0.764**	0.165
Overall F ($\lambda = 8$)	6.286	5.935	4.570	**8.749**	3.984

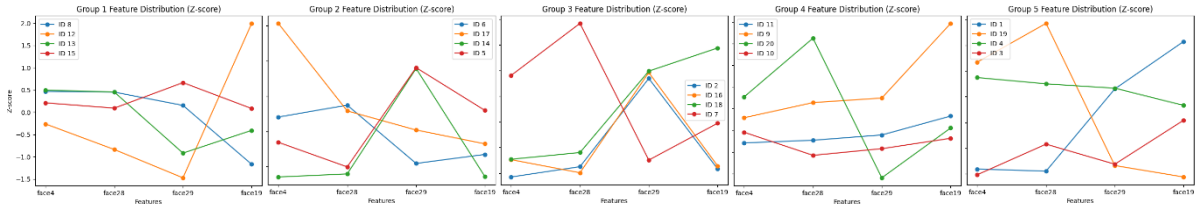

Figure 1. Group member alignment and aggregated intensity distributions.

2.3 Pairwise analysis of temporal activation patterns

Beyond group formation, proactive support can further inform subsequent instructional design. Through group awareness tools, multimodal indicators can be visualized before and during group learning activities. At the intensity level, detected emotional constructs such as boredom may be used to optimize role assignment, thereby facilitating balanced engagement among members. At the temporal structure level, activation patterns can be interpreted in connection with the structure of the learning materials. Confusion points may indicate opportunities for peer explanation and discussion.

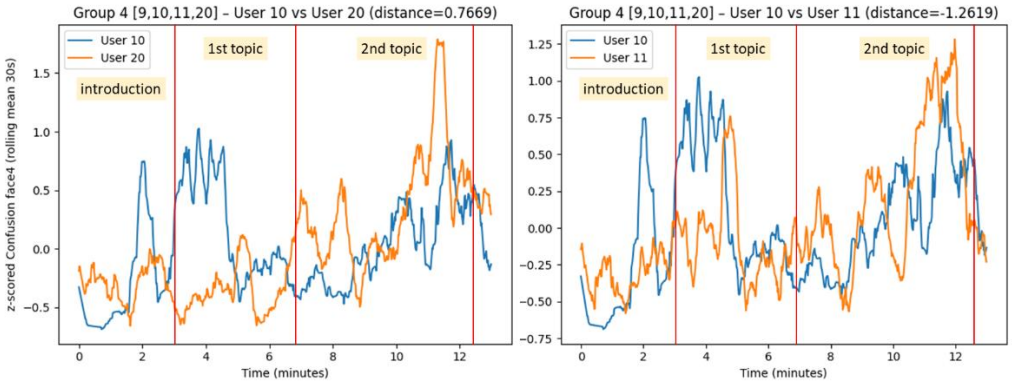

Figure 2. Comparison of temporal activation patterns between pairs.

Taking Group 4 with the highest *F(pairwise)* as an example, Learners 10 and 20 exhibit compensatory patterns with a relatively large pairwise distance. Divergence appears at the start of the first topic and again at the start of the second topic (see Figure 2), suggesting need for mutual assistance according to ZPD theory. In contrast, Learners 10 and 11 shared highly similar activation trajectories, with common confusion concentrated at the commence of the first topic and in the latter part of the second topic. Such synchronized confusion points indicate shared conceptual challenges. In this case, revisiting those segments through direct explanation from teachers may be more appropriate than relying solely on group discussion.

3. Discussion and Conclusion

This work elucidates the practical application of multimodal data to support group learning design. By integrating genetic algorithms with pairwise indicators, different modalities are fused at both the intensity level and the structural level, enabling interpretable insights for practitioners. As group awareness tools, the visualizations of these multimodal indicators hold potential to guide group learning processes. Teachers can even use this data to simulate possible interaction patterns before the actual group activities, allowing early detection of potential issues and more deliberate planning (Yan et al., 2025). Unlike conventional research streams on ongoing group processes and outcomes, this study adopts a data-driven paradigm in which antecedent learning data inform group design and prediction (Liang et al., 2024).

Future work should validate the proposed group formation strategy in authentic collaborative settings, scrutinizing its impact on subsequent learning performance. Expanding the framework to include broader modalities across diverse task contexts remains a forward-looking agenda as well. Moreover, employing generative AI to recommend adaptive grouping strategies based on various learning contexts and to assist in interpreting group awareness indicators could further broaden the applicability of this design.

Acknowledgements

This work is partly supported by JSPS KAKENHI Grant Number 25K21357 and JST CSTI SIP Program Grant Number JPJ012347.

References

Cohn, J. F., Ambadar, Z., & Ekman, P. (2007). Observer-based measurement of facial expression with the Facial Action Coding System. *The handbook of emotion elicitation and assessment*, *1*(3), 203-221.

Daza, R., Shengkai, L., Morales, A., Fierrez, J., & Nagao, K. (2025, October). SMARTe-VR: Student monitoring and adaptive response technology for e-learning in virtual reality. In Proceedings of the International Workshop on Intelligent Immersification in the Metaverse: AI-Driven Immersive Multimedia (pp. 15-24).

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., & Keogh, E. J. (2008). Querying and mining of time series data: Experimental comparison of representations and distance measures. *Proc. VLDB Endow.*, *1*(2), 1542-1552.

Giannakos, M., & Cukurova, M. (2023). The role of learning theory in multimodal learning analytics. *British Journal of Educational Technology*, *54*(5), 1246-1267.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., ... & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and education: artificial intelligence*, *3*, 100074.

Kim, J. M., Ng, P. T. Y., Pinto, N. K., Lai, K. C. H., Wu, E. Y. T., Ngan, O. M. Y., ... & Tang, F. M. K. (2026, January). New Concept of Digital Learning Space for Health Professional Students: Quantitative Research Analysis on Perceptions. In *Informatics* (Vol. 13, No. 1, p. 13). MDPI.

Liang, C., Majumdar, R., Horikoshi, I., & Ogata, H. (2024). Data-driven support infrastructure for iterative team-based learning. *IEEE Access*, *12*, 65967-65980.

Liang, C., Takii, K., & Ogata, H. (2025, September). Pairwise learner model for collaborative learning and its application in genetic group formation. In *International Conference on Learning Evidence and Analytics*.

Reimann, P., & Baker, M. J. (2025). Editorial notes: Mechanisms as a unifying construct for CSCL research. *International Journal of Computer-Supported Collaborative Learning*, *20*(4), 415-424.

Saqr, M. (2024). Group-level analysis of engagement poorly reflects individual students' processes: Why we need idiographic learning analytics. *Computers in Human Behavior*, *150*, 107991.

Tao, L., Song, Y., & Fu, J. (2025). Exploring students' self-regulated learning behavioural patterns and perceptions in an English speaking task within a generative AI-supported immersive VR. *Computers & Education*, 105515.

Yan, Y., Liang, C., & Ogata, H. (2025, November). Simulating Collaborative Learning with Data-Driven LLM-Agents. In *International Conference on Collaboration Technologies and Social Computing* (pp. 135-143).