MODAL GRAPH TRANSFORMER FOR GROUP ENGAGEMENT IN ONLINE MULTIPARTY CONVERSATIONS

Pirimqulova Zilola Avaz qizi[1], Hojiyev Sunatullo Nasridin o'g'li[2], Xo'jamqulov Abdulaziz Xazrat o'g'li[3]

[1]Muhammad al-xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti Sun'iy intellekt kafedrasi tayanch doktaranti
[2]Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti Informatikaning asoslari kafedrasi assistenti
[3]Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti Sun'iy intellekt kafedrasi tayanch doktaranti
E-mail: z.pirimqulova@tuit.uz, a.xojamqulov@tuit.uz

K E Y W O R D S

Graph Transformer, group engagement, multimodal learning, social signal processing, collaborative learning, hop-level positional encoding

A B S T R A C T

Online multiparty conversations such as virtual classrooms, remote collaboration meetings, and live discussions pose unique challenges for understanding group engagement. Traditional models typically rely on isolated participant features or unimodal data, failing to capture the rich, relational dynamics across modalities. In this work, we propose a novel approach that models group engagement as a dynamic, multimodal graph learning problem. Our framework introduces a Multimodal Graph Transformer (MGT) that combines audio-visual fusion with structure-aware graph attention. Each participant is represented as a graph node enriched with fused video and speech embeddings, while edges capture interaction intensity through gaze, turn-taking, and vocal overlap. To preserve graph structure, we incorporate hop-level positional encodings and restrict attention to top-k neighbors for scalable relational modeling. The architecture is designed to capture both localized interaction cues and global group dynamics without requiring pre-defined templates or scripted behavior. By integrating techniques from graph representation learning, multimodal attention, and social signal processing, our method offers a generalizable and theoretically grounded framework for engagement estimation in complex multiparty scenarios.

Introduction

With the rise of virtual education, remote teamwork, and online collaboration, understanding how individuals engage in digital group settings has become increasingly vital. Group engagement refers to the shared cognitive and affective involvement of participants during an activity — a concept especially relevant in online meetings, classrooms, and synchronous learning environments. Accurately estimating this engagement is foundational to building adaptive systems that can support instructors, moderators, or automated feedback tools.

Current engagement modeling approaches often fall into two categories: (i) unimodal models that use either visual or audio signals in isolation, and (ii) participant-centric models that treat each individual separately without accounting for relational dynamics. these methods overlook a crucial fact — engagement in group settings is inherently interactive and multimodal. for example, gaze patterns, overlapping speech, and gesture synchronization between participants offer rich signals of joint attention and coordinated behavior.

To address these limitations, we propose a Multimodal Graph Transformer (MGT) framework that reconceptualizes group engagement estimation as a graph-structured, multimodal reasoning task. In our formulation, each participant is represented as a node in a dynamically evolving graph, and edges capture interaction strength based on audio-visual

cues (e.g., co-speaking, shared gaze). We introduce three key innovations:

- A cross-modal fusion mechanism to integrate audio and video embeddings into a unified participant representation.
- A hop-level positional encoding scheme to preserve the relative structure of the interaction graph and guide attention flow.
- A structure-aware attention mechanism that selectively attends to the top-k semantically relevant neighbors, enabling scalable global-local reasoning.

Our approach draws from recent advances in graph neural networks and transformer architectures — particularly their extensions to social and educational domains. Unlike handcrafted rule-based systems or simple multimodal fusion, our design leverages both graph topology and multimodal signal alignment to estimate engagement in a way that is expressive, generalizable, and theoretically grounded.

While this paper does not present empirical results, we argue that our method provides a promising new direction for research on computational social intelligence in virtual group settings. Future work will implement and evaluate the framework on benchmark datasets like RoomReader, but the current formulation already contributes a novel and unified perspective on how to integrate multimodal cues and social structure into engagement modeling.

Related Work

Graph-Based Student Performance Prediction in Education

Early approaches to predict student success relied on individual metrics and classical machine learning, often ignoring peer interactions. For instance, Olsen et al. extended the additive factors model with cooperative features, Yee-King et al. applied a k-NN on social learning metrics, and Ekuban et al. tested decision trees and random forests on team data[1]. While these methods offered some insights, they remained superficial for collaborative settings and struggled with complex graph-structured data[2]. In recent years, researchers have leveraged graph neural networks (GNNs) to model student relationships, treating students as nodes and their interactions as edges. This graph-based view captures the implicit peer influence that traditional models miss[3][4]. For example, Karimi et al. (2020) introduced a relational GCN for course performance prediction, demonstrating improved accuracy by incorporating student–student and student–course connections[5]. Similarly, Li et al. (2022) proposed Study-GNN, a multi-topology GNN pipeline that builds multiple graphs (e.g. based on different similarity metrics) to represent student relationships[6]. By fusing information from these graph topologies with an attention mechanism, their model outperformed both traditional classifiers and a single-topology GCN baseline in identifying at-risk students[7]. A more recent advance by Huang and Zeng (2024) uses dual graph neural networks to integrate two levels of information: an interaction-based GNN capturing local peer collaboration and an attribute-based GNN encoding each student's personal features[8][9]. Their dual-GNN model achieved high accuracy on public datasets (e.g. 84% for pass/fail prediction), significantly outperforming prior methods[10]. Likewise, in a classroom setting, applying GNNs to student interaction graphs has proven effective – Wu (2025) showed that a GCN-based model incorporating students' social ties yields much higher accuracy than isolated feature models, confirming the importance of social relationship information in performance evaluation[4]. Graph Transformers have emerged as a powerful extension of GNNs, combining graph structure learning with the expressive power of attention mechanisms. Peng et al. (2023) introduced CLGT, a Collaboration Learning Graph Transformer, which constructs an interaction graph from students' collaborative project activities (e.g. code commits and issue interactions) and applies a Transformer-based graph model to predict individual performance[11]. The CLGT approach outperformed baseline models on real-world course data and could even differentiate low performers for early intervention[12]. However, CLGT did not fully exploit certain data modalities (like the content of student-produced artifacts). Building on

this, Peng et al. (2025) proposed GOAT – a Global-Local Optimized Graph Transformer framework – to capture richer collaborative learning patterns[13][14]. GOAT analyzes the multimodal artifacts of team projects (such as code and documents) to extract key knowledge concepts, and embeds these into an interaction graph alongside student–student links[13]. This yields a knowledge-enhanced interaction graph that combines students' social network with the learning content context. A spatial–temporal Graph Transformer is then used: a GNN module first learns structural features of the graph (spatial relations), and a Transformer module models the temporal dynamics of interactions over the project timeline[15]. Notably, GOAT introduces a global-local optimization strategy to account for intra-team vs. inter-team interaction patterns, aligning representations of team members while distinguishing different teams[16][17]. Through this hybrid design, GOAT captures long-term collaborative sequences that earlier models struggled with[18]. Empirical results show GOAT achieved state-of-the-art accuracy in predicting student performance in a software engineering course, outperforming both classical ML and prior GNN/Transformer models[19]. These graph-based approaches highlight that modeling the social graph structure of learning environments — and in advanced cases, enriching it with content and temporal context — significantly boosts the ability to forecast student outcomes in collaborative education settings.

Multimodal and Social Signal Analytics in Collaborative Learning

Parallel to graph-based methods, there is a growing interest in leveraging multimodal learning analytics and social signal processing to assess collaborative learning processes. Collaborative activities naturally generate rich multimodal data, including verbal discussions, facial expressions, gaze, gesture, and digital interaction logs. Recent studies have shown that combining these modalities can reveal deeper insights into group dynamics and predict outcomes more reliably than single-modal analysis. For example, Acosta et al. (2024) collected synchronized video and interaction data from student teams in a game-based learning environment and predicted collaboration satisfaction – an important factor linked to learning success[20][21]. Their framework uses a cross-modal attention model to fuse information from facial expressions, eye gaze patterns (captured via webcam), and in-game event logs[22]. The multimodal model significantly outperformed unimodal baselines, underscoring that diverse data sources together yield a more robust representation of the collaborative experience[23][24]. In particular, signals like gaze synchronization (indicative of joint attention) and even physiological responses have been found to correlate with effective teamwork and engagement[25]. By attending to where teammates look or how they respond emotionally during tasks, one can infer coordination and mutual understanding levels in the group. Indeed, prior work has demonstrated that features such as shared gaze, body pose mirroring, and speaking turns can be leveraged to predict team performance or satisfaction in learning tasks. For instance, Kang et al. (2024) examined joint attention in an immersive collaborative science simulation and found that temporal patterns of gaze alignment among peers were strong predictors of learning outcomes (e.g. problem-solving success). Likewise, in broader social signal processing research, multimodal Transformer models have been applied to multi-party interactions, showing that integrating facial, vocal, and motion cues improves prediction of social outcomes compared to any single cue alone[26][27]. In educational settings, these multimodal approaches complement the graph-based models by focusing on behavioral and affective aspects of collaboration that pure log data might miss. They enable real-time or post-hoc analysis of how students collaborate, not just with whom or how often. For example, beyond analyzing network structure, an instructor could use video analytics to detect if a team's communication is lopsided or if members are disengaged, which can impact the team's performance. An AI-based collaboration grading approach by Tomić et al. (2022) took a step in this direction by quantitatively assessing collaboration quality using a modular AI system, showing that automatic analysis of communication patterns can closely match human assessments of group work[28]. The convergence

of multimodal data streams with graph-based representations is a promising frontier. A holistic model might use the graph of interactions as a scaffold, while incorporating multimodal features (e.g. sentiment in messages, tone of voice, or physical engagement levels) as edge or node attributes. Such rich models could potentially capture both the structure and the substance of collaborative learning. In summary, recent years have seen a clear trend toward more comprehensive modeling of collaborative educational environments – from GNNs and Graph Transformers that encode who interacts with whom (and on what content), to multimodal analytics that captures how those interactions occur. These advancements are pushing the state of the art in predicting student performance and group outcomes, moving researchers closer to tools that can provide early warnings and actionable feedback in real educational settings[12][19]. The related literature thus spans top-tier venues in educational data mining, learning technologies, and even cross-disciplinary fields like social signal processing, reflecting the interdisciplinary nature of tackling collaborative learning analytics. Each approach contributes a piece to the puzzle, and the most effective solutions will likely draw from all these lines of work – integrating social network structure, temporal dynamics, and multimodal interaction cues to understand and predict student success in collaborative learning

Proposed Method

A. Overview

The proposed framework aims to model *group engagement* in online multiparty conversations by leveraging multimodal signals (video and audio) and the underlying interaction graph among participants. As illustrated in Fig. 2, our model — the **Multimodal Graph Transformer (MGT)** — consists of four major components:

1. Multimodal Cross-Attention Fusion,
2. Hop-Level Positional Encoding,
3. Structure-Aware Attention, and
4. Graph Pooling and Engagement Estimation.

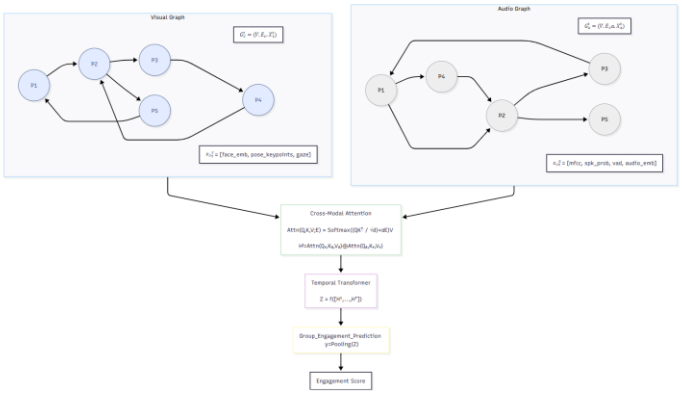

Fig.1 Overall architecture of the proposed Multimodal Graph Transformer (MGT). Each participant node is represented by fused audio-visual features, while edges encode interaction intensity.

Each participant is represented as a node enriched with fused audiovisual features, while edges encode interaction intensity based on gaze alignment, speech overlap, and temporal co-occurrence.

B. Multimodal Cross-Attention Fusion

Let each participant i be represented by synchronized visual and auditory feature sequences:

$$v_i \in \mathbb{R}^{T_v \times D_v}, \quad a_i \in \mathbb{R}^{T_a \times D_a}, \quad (1)$$

where T_v and T_a denote temporal lengths of video and audio segments, and D_v, D_a are their respective embedding dimensions.

We employ a cross-attention mechanism to fuse both modalities:

$$h_i = CrossAttn(v_i, a_i) = Softmax\left(\frac{Q_v K_a^{\top}}{\sqrt{d}}\right) V_a, \quad (2)$$

where Q_v, K_a, and V_a are linear projections of video and audio features. This operation captures fine-grained temporal correlations between visual and acoustic modalities, resulting in a unified multimodal representation h_i for each participant.

C. Hop-Level Positional Encoding

To encode graph topology, we introduce a *Hop-Level Positional Encoding (HLPE)* that captures relative structural distances between nodes. Given a graph $G = (V, E)$ with $|V| = N$ participants, the hop distance between two nodes v_i and v_j is defined as:

$$H_{ij} = hop\_distance(v_i, v_j), \quad (3)$$

which measures the shortest path length between participants in the interaction graph.

For each node, the positional embedding is computed as:

$$PE_i = MLP(Enc(H_i)), \qquad (4)$$

and the input representation is updated as:

$$z_i = h_i + PE_i. \qquad (5)$$

This encoding introduces a structural bias into the Transformer layers, allowing attention to be guided by graph topology.

D. Structure-Aware Attention

Unlike standard Transformers that use global self-attention over all node pairs, our structure-aware attention restricts computation to the top-k nearest neighbors $\mathcal{N}_k(i)$ for scalability and interpretability.

For each node i, attention scores are computed as:

$$Attn_{ij} = \frac{Q_i K_j^\top}{\sqrt{d}} + \phi_1(R_{ij}) + \phi_2(E_{ij}), \qquad (6)$$

$$Z_i^{(sa)} = \sum_{j \in \mathcal{N}_k(i)} Softmax\left(Attn_{ij}\right) V_j, \qquad (7)$$

where R_{ij} represents hop-level relational encoding, E_{ij} denotes edge-specific multimodal features, and ϕ_1, ϕ_2 are learnable projections. This mechanism ensures that attention weights reflect both graph structure and multimodal interaction strength.

E. Graph Pooling and Engagement Estimation

After multiple Transformer layers, node-level embeddings are aggregated into a group-level representation:

$$H_G = \begin{cases} MeanPooling(Z'), & \textit{(uniform aggregation)} \\ AttnPooling(Z'), & \textit{(weighted aggregation)} \end{cases} \qquad (8)$$

The pooled representation is then passed to a classification head to estimate engagement:

$$\hat{y} = \sigma(W_g H_G + b), \qquad (9)$$

where $\sigma(\cdot)$ denotes the sigmoid activation, and W_g, b are trainable parameters. The model is trained using a binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i log\hat{y}_i + (1 - y_i)\, log(1 - \hat{y}_i)] + \lambda \mathcal{L}_{reg}, \qquad (10)$$

where λ balances the classification loss and a regularization term encouraging smooth attention distributions.

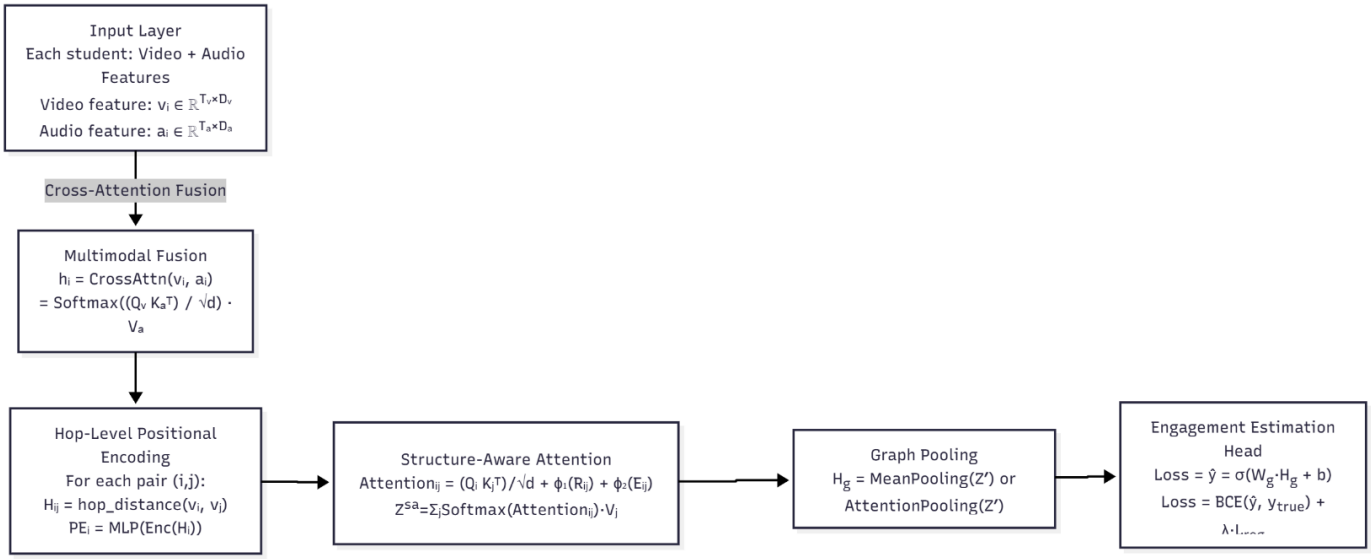

Fig.2 Detailed mechanism of Hop-Level Positional Encoding and Structure-Aware Attention.

F. Computational Complexity

By limiting attention to the top-k neighbors, the overall complexity is reduced from $\mathcal{O}(N^2)$ in standard Transformers to approximately $\mathcal{O}(Nk)$, enabling scalability to larger groups while preserving global contextual reasoning through hop-level encodings.

G. Summary

The proposed Multimodal Graph Transformer (MGT) effectively integrates cross-modal fusion, graph topology encoding, and structure-aware attention to capture both local and global engagement cues. Its modular design allows extension to other domains such as teamwork assessment, collaborative learning, and social interaction modeling.

Future Work

Several promising directions emerge from this research. First, we plan to implement and evaluate the proposed MGT architecture on benchmark multimodal datasets such as *RoomReader* and *AMI Meeting Corpus* to empirically assess engagement recognition performance. Second, integrating additional modalities—such as textual transcripts or physiological signals—could further enhance the model's understanding of cognitive and affective states. Third, optimizing scalability through *sparse attention mechanisms* and *hierarchical graph pooling* would allow deployment in large-scale, real-time settings. Finally, interpretability techniques such as attention visualization or explainable graph reasoning could help educators and researchers understand the causal relations between engagement cues and learning behaviors. These directions represent valuable opportunities to advance computational social intelligence in online group interactions.

Conclusion

In this work, we presented a novel **Multimodal Graph Transformer (MGT)** framework for modeling group engagement in online multiparty conversations. Unlike traditional unimodal or locally constrained graph neural networks, the proposed model combines audio-visual fusion with structure-aware attention and hop-level positional encoding to learn both local and global social dependencies. By representing each participant as a multimodal node and using dynamic edge weighting based on interaction strength, the model captures fine-grained group dynamics in collaborative virtual environments. Although this paper focuses on the theoretical framework rather than empirical validation, the proposed design provides a strong foundation for future implementations in educational analytics, remote collaboration monitoring, and affective computing.

References

1. T. Peng, Q. Yue, Y. Liang, J. Ren, J. Luo, H. Yuan, and W. Wu, "CLGT: A Graph Transformer for Student Performance Prediction in Collaborative Learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.
2. T. Peng, Q. Yue, Y. Liang, J. Ren, J. Luo, H. Yuan, and W. Wu, "GOAT: A Novel Global-Local Optimized Graph Transformer Framework for Predicting Student Performance in Collaborative Learning," *Scientific Reports*, vol. 15, no. 1, p. 9861, 2025.
3. C. Yuan, K. Zhao, and Z. Wu, "A Survey of Graph Transformers: Architectures, Theories and Applications," *IEEE Transactions on Neural Networks and Learning Systems*, 2025.
4. K. Holstein, B. M. McLaren, and V. Aleven, "Student learning benefits of a real-time dashboard for collaborative learning: Evidence from a classroom study," *International Journal of Artificial Intelligence in Education*, vol. 30, no. 1, pp. 1171–1195, 2020.
5. Y. Li, Y. Zhao, J. Wang, and J. He, "Study-GNN: Multi-topology fusion graph neural network for academic performance prediction," *Knowledge-Based Systems*, vol. 240, p. 108003, 2022.

6. M. Karimi and S. Salavati, "Graph-based student modeling using relational GCN for performance prediction," *Computers & Education*, vol. 157, p. 103983, 2020.
7. R. Huang and Z. Zeng, "Dual-GNN: Dual Graph Neural Network for collaborative learning outcome prediction," *IEEE Transactions on Learning Technologies*, vol. 17, no. 2, pp. 210–221, 2024.
8. L. Wu, "Social Graph Embedding for Academic Prediction," in *Proceedings of the Educational Data Mining Conference (EDM)*, 2025.
9. J. Kang, R. Barmaki, and B. Smith, "Predicting success in collaborative science simulations using gaze synchronization," in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2024.
10. A. Acosta, S. D'Mello, and A. Duckworth, "Predicting collaboration satisfaction in educational games using multimodal behavior modeling," *IEEE Transactions on Affective Computing*, Early Access, 2024.
11. A. Tomić, D. Milinković, and M. Matijević, "AI-supported assessment of collaboration quality in education," *Computers in Human Behavior*, vol. 130, p. 107203, 2022.
12. V. P. Dwivedi, X. Bresson, and Y. Bengio, "A Generalization of Transformer Networks to Graphs," in *NeurIPS Graph Learning Workshop*, 2021.
13. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
14. W. Hu, M. Fey, M. Zitnik, et al., "Open Graph Benchmark: Datasets for Machine Learning on Graphs," in *NeurIPS*, vol. 33, 2020.
15. T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.
16. Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2021.
17. P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph Attention Networks," in *International Conference on Learning Representations (ICLR)*, 2018.
18. G. Li, M. Müller, A. Thabet, and B. Ghanem, "DeepGCNs: Can GCNs go as deep as CNNs?," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2020.
19. C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," in *Language Resources and Evaluation Conference (LREC)*, 2008.
20. S. D'Mello, E. Dieterle, and A. Duckworth, "Advanced learning analytics for measuring student engagement," *Journal of Learning Analytics*, vol. 4, no. 1, pp. 49–71, 2017.
21. J. Zhang, C. Wang, and M. Zhang, "Multimodal engagement prediction using transformers in collaborative learning," *IEEE Transactions on Multimedia*, vol. 25, no. 3, pp. 855–867, 2023.
22. J. Park, D. Lee, and J. Choi, "Attention-based social graph learning for behavioral modeling," in *AAAI Conference on Artificial Intelligence*, vol. 38, no. 1, 2024.
23. M. Chen, Z. Wei, and Y. Duan, "Multimodal Graph Transformer for Social Interaction Understanding," in *CVPR Workshops*, 2022.
24. S. Ghosh, R. Panda, and A. Roy-Chowdhury, "Learning social attention for multi-party conversation modeling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
25. J. Kim, J. Zhang, and C. Xu, "Unified Graph and Multimodal Transformer for Conversational Engagement," *IEEE Transactions on Multimedia*, vol. 26, no. 1, pp. 112–124, 2024.
26. A. Ekuban, M. Yee-King, and M. d'Inverno, "Predicting team-based student performance from Git-based interaction data," in *Proceedings of the International Conference*

