Cross-modal Affinity-aligned Multimodal Learning Analytics for Predicting Student Collaboration Satisfaction in Game-Based Learning

Anonymous CVPR submission

Paper ID 9

Abstract

Collaborative game-based learning environments offer rich opportunities for small-group knowledge construction, yet automatically predicting student collaboration satisfaction remains challenging. A critical barrier is modality degradation: in educational deployments, individual modalities such as eye gaze exhibit inconsistent informativeness across student cohorts, causing implicit attention-based fusion to produce brittle multimodal representations. We propose the Affinity-Aligned Multimodal Learning Analytics (AAMLA) framework, whose core contribution is the Cross-modal Affinity-guided Modality Alignment (CAMA) module, which explicitly models inter-modal relationships via affinity matrices and enforces cross-modal consistency through contrastive learning, enabling adaptive suppression of uninformative modalities without discarding them. AAMLA further applies modality-specific projection layers to map heterogeneous features — facial action units, head pose, eye gaze, and interaction trace logs — into a unified semantic space prior to alignment. Experiments on 50 middle school students in the EcoJourneys collaborative learning environment demonstrate consistent improvements over unimodal baselines and prior cross-attention approaches under standard and modality degradation conditions, with SHAP and t-SNE analyses confirming that CAMA produces robust, interpretable cross-modal representations for student collaboration modeling.

1. Introduction

Collaborative game-based learning environments provide students with immersive opportunities for small-group problem solving, knowledge construction, and collaborative inquiry [15, 20, 51, 52]. Within these environments, student satisfaction with collaborative experiences plays a pivotal role in shaping teamwork dynamics, sustaining motivation, and ultimately influencing learning outcomes [51]. Accurately gauging collaboration satisfaction can offer actionable insights for instructors and intelligent tutoring systems to provide timely, targeted interventions that support productive group interactions. However, existing work predominantly relies on post-hoc survey analysis [2, 25, 39], leaving automated assessment methods relatively underexplored.

Multimodal behavioral signals — including facial action units, head pose, eye gaze, and interaction trace logs — offer rich and complementary cues for understanding student collaboration dynamics [13, 37]. However, a fundamental challenge in educational deployments is that individual modalities are frequently unreliable in practice: sensor noise, occlusion, and student behavioral variation cause modality quality to fluctuate across students and sessions [30, 45]. This problem is especially pronounced for gaze features, which exhibit inconsistent informativeness across student cohorts [18, 19] — a form of modality degradation that cannot be addressed by simply discarding the affected modality, as it may still carry useful signals in other contexts. Despite these challenges, existing multimodal approaches in learning analytics rely on implicit cross-attention fusion mechanisms [27] that assume all modalities contribute equally informative signals, and lack explicit mechanisms to handle modality degradation gracefully. When a weak or noisy modality dominates attention, cascading errors propagate through the entire fusion pipeline, degrading prediction robustness across student cohorts.

To address these limitations, we propose the **Affinity-Aligned Multimodal Learning Analytics (AAMLA)** framework for predicting student collaboration satisfaction in collaborative game-based learning. The core contribution of AAMLA is the **Cross-modal Affinity-guided Modality Alignment (CAMA)** module, which explicitly models inter-modal relationships through affinity matrices and enforces cross-modal consistency via contrastive learning [10, 46]. Unlike implicit fusion, CAMA learns to adaptively suppress uninformative or degraded modalities — such as inconsistent gaze signals — without discarding them entirely, ensuring stable multimodal representations

even under severe modality degradation. We further apply modality-specific linear projection layers to map heterogeneous features — facial action units [4], head pose, gaze, and BERT-based trace log embeddings [27] — into a unified semantic space prior to alignment, enabling direct cross-modal comparison.

Our research addresses the following questions:

- **RQ1:** What are the individual modalities that contribute most significantly to predicting student collaboration satisfaction [6, 40, 41]?
- **RQ2:** How does CAMA improve robustness over implicit cross-attention fusion under modality degradation conditions including gaze dropout, feature perturbation, and full modality absence?
- **RQ3:** What modality combinations are most effective, and what do affinity matrix visualizations and t-SNE distributions reveal about cross-modal alignment quality under degradation?

We conduct experiments on multimodal data collected from 50 middle school students interacting with the EcoJourneys collaborative game-based learning environment [8]. Results demonstrate that AAMLA consistently outperforms unimodal baselines and the prior cross-attention approach [1] across all modality combinations, and maintains robust performance under three categories of modality degradation. t-SNE visualizations [43] and affinity matrix analyses confirm that CAMA produces semantically coherent cross-modal representations, offering interpretable insights into the multimodal behavioral patterns most predictive of student collaboration satisfaction.

2. Related Work

Collaborative Game-Based Learning. Collaborative game-based learning environments offer significant promise in fostering collaborative skills and providing insights into student dynamics [8, 11, 26, 28]. Digital games serve as a conduit to explore the relationship between learning outcomes and collaborative gameplay [3, 36]. However, a systematic framework elucidating the elements of collaborative learning within these environments remains elusive [44], while the pivotal importance of student satisfaction and motivation in educational tool design has been consistently highlighted [15].

Student Satisfaction. Prior work has investigated the relationship between student satisfaction and learning processes across diverse settings [2, 25, 39, 50]. Collaborative engagement and instructional quality emerge as pivotal determinants of satisfaction, which is consistently associated with higher learning outcomes [49]. However, existing research predominantly relies on post-hoc surveys, overlooking automated assessment methodologies. The closest prior work [32] predicts peer satisfaction in dyadic settings from multimodal cues, yet extending this to non-dyadic small groups remains an open challenge.

Multimodal Learning Analytics. Multimodal learning analytics provides rich indicators of collaborative engagement [6, 31, 40, 41], with prior work demonstrating effectiveness in predicting affective states during collaborative activities [12], modeling cognitive load from physiological signals [7], and identifying collaboration quality from speech and non-verbal cues [35]. Griffith et al. [17] further demonstrate how co-creative dialogue states influence partner satisfaction, while temporal integration of multimodal features has been shown to improve prediction of learning gains [34]. Despite these advances, existing frameworks rely on implicit fusion mechanisms that treat all modalities equally, without accounting for the varying informativeness of individual modalities across different student cohorts and interaction contexts. The exploration of collaboration satisfaction prediction in game-based learning environments, particularly under such modality imbalance, remains underexplored.

Modality Degradation and Robustness. A fundamental challenge in multimodal learning is that individual modalities may be unavailable, corrupted, or simply uninformative in practice [30, 45]. In educational settings, this problem is especially pronounced: sensor noise, occlusion, and student behavioral variation can render certain modalities unreliable across sessions or cohorts. Prior work has explored missing modality scenarios through shared-specific feature modeling [45] and Bayesian approaches for severely missing modalities [30], while test-time adaptation methods have been proposed to handle modality reliability bias at inference [48]. In the context of deepfake detection, Le and Woo [23] demonstrate that quality-agnostic learning can improve robustness against compression-induced feature degradation, and UMCL [22] further shows that generating complementary modalities from a single source mitigates unequal modality degradation. However, these approaches are not designed for the educational domain, where modality degradation is behavioral rather than compression-induced. In our setting, gaze features exhibit inconsistent informativeness across student cohorts [18, 19], motivating an explicit mechanism to suppress uninformative modalities dynamically rather than discarding them entirely.

Feature Alignment in Multimodal Learning. Feature alignment is fundamental to robust multimodal fusion, ensuring that modality-specific representations occupy a coherent shared semantic space [16, 21, 47]. Contrastive learning approaches reduce modality gaps through explicit cross-modal regularization [10, 29, 46], while affinity-based methods model global and local inter-modal relationships to improve interpretability and robustness [24]. Recent work in multimodal deepfake detection [22] demonstrates that explicit alignment of self-generated modalities substantially outperforms implicit attention-based fusion under de-

graded conditions. Nevertheless, existing alignment methods largely assume that all modalities contribute meaningful signal — an assumption that fails when weak modalities introduce noise into the fusion process. Our proposed CAMA strategy extends affinity-based alignment to explicitly down-weight uninformative modalities, ensuring stable cross-modal coordination even under behavioral modality degradation in collaborative learning settings.

3. Methodology

3.1. Framework Overview

In this section, we present the AAMLA framework for collaboration satisfaction prediction. As illustrated in Fig. 1, AAMLA processes four complementary modality streams from student interaction data — facial action units (AU), head pose, eye gaze, and trace logs — through modality-specific encoders. The resulting embeddings are unified via learnable projection layers and aligned by the proposed *Cross-modal Affinity-guided Modality Alignment* (CAMA) module, which explicitly models inter-modal relationships through affinity matrices to suppress uninformative modalities and ensure semantic consistency across heterogeneous feature spaces.

3.2. Multimodal Feature Encoding

Our framework operates on four complementary modalities extracted from student interactions within the EcoJourneys learning environment [8]. Each modality captures distinct aspects of collaborative behavior that contribute uniquely to satisfaction prediction.

Facial Action Units. Facial action units (AUs) capture the contractions and relaxations of facial muscles, providing direct cues about students' emotional states and engagement during collaboration [5, 33]. We extract 17 AU intensity values per frame using OpenFace 2.0 [4], producing a sequence $\mathbf{X}_{\text{AU}} \in \mathbb{R}^{T\times 17}$ over T timesteps. The AU encoder processes this sequence through a GRU:

$$\mathbf{e}_{\text{AU}} = \Phi_{\text{AU}}(\mathbf{X}_{\text{AU}}), \tag{1}$$

where Φ_{AU} represents the GRU-based AU encoder and $\mathbf{e}_{\text{AU}} \in \mathbb{R}^{1\times d_{\text{AU}}}$ is the resulting temporal embedding.

Head Pose. Head pose information captures students' head location and orientation relative to the camera, providing cues about attentiveness and engagement during collaborative activities [9, 42]. We extract 6 pose features per frame (3D translation and rotation coordinates) using OpenFace 2.0 [4], producing $\mathbf{X}_{\text{pose}} \in \mathbb{R}^{T\times 6}$:

$$\mathbf{e}_{\text{pose}} = \Phi_{\text{pose}}(\mathbf{X}_{\text{pose}}), \tag{2}$$

where Φ_{pose} represents the GRU-based pose encoder.

Eye Gaze. Eye gaze features detail the direction each eye is looking relative to the camera, capturing visual attention patterns during collaboration [18, 19, 38]. We extract 6 gaze direction features per frame (3D coordinates per eye) using OpenFace 2.0 [4], producing $\mathbf{X}_{\text{gaze}} \in \mathbb{R}^{T\times 6}$:

$$\mathbf{e}_{\text{gaze}} = \Phi_{\text{gaze}}(\mathbf{X}_{\text{gaze}}), \tag{3}$$

where Φ_{gaze} represents the GRU-based gaze encoder. As demonstrated in prior work [32] and our preliminary analysis, gaze features exhibit inconsistent informativeness across student cohorts, motivating our explicit alignment mechanism to suppress their contribution dynamically when uninformative.

Interaction Trace Logs. Trace logs document students' in-game actions, including NPC interactions, locations visited, evidence collected, and chat messages, providing rich contextual information about collaborative behavior [17, 34]. Each trace event is encoded as a 768-dimensional embedding using a BERT-based sentence encoder fine-tuned on in-game text via unsupervised pre-training [27]. The trace encoder projects these embeddings through a feedforward network:

$$\mathbf{e}_{\text{trace}} = \Phi_{\text{trace}}(\mathbf{X}_{\text{trace}}), \tag{4}$$

where Φ_{trace} represents the feedforward trace encoder and $\mathbf{X}_{\text{trace}} \in \mathbb{R}^{T\times 768}$.

Feature Projection. After encoding, a key challenge is ensuring consistent feature distributions across modalities with heterogeneous dimensionalities. We apply modality-specific learnable linear projection layers to map all features into a unified d-dimensional semantic space:

$$\hat{\mathbf{e}}_{\text{AU}} = \mathbf{W}_{\text{AU}}\mathbf{e}_{\text{AU}} + \mathbf{B}_{\text{AU}}, \quad \hat{\mathbf{e}}_{\text{pose}} = \mathbf{W}_{\text{pose}}\mathbf{e}_{\text{pose}} + \mathbf{B}_{\text{pose}},$$
$$\hat{\mathbf{e}}_{\text{gaze}} = \mathbf{W}_{\text{gaze}}\mathbf{e}_{\text{gaze}} + \mathbf{B}_{\text{gaze}}, \quad \hat{\mathbf{e}}_{\text{trace}} = \mathbf{W}_{\text{trace}}\mathbf{e}_{\text{trace}} + \mathbf{B}_{\text{trace}}, \tag{5}$$

where $\mathbf{W}_{\{\cdot\}}$ and $\mathbf{B}_{\{\cdot\}}$ are modality-specific learnable parameters, and all projected features $\hat{\mathbf{e}}_{\{\cdot\}} \in \mathbb{R}^{1\times d}$ share a common dimensionality $d = 128$.

The projected features are concatenated to form a unified multimodal representation:

$$\mathbf{U} = [\hat{\mathbf{e}}_{\text{AU}}, \hat{\mathbf{e}}_{\text{pose}}, \hat{\mathbf{e}}_{\text{gaze}}, \hat{\mathbf{e}}_{\text{trace}}] \in \mathbb{R}^{m\times d}, \tag{6}$$

where $m = 4$ represents the number of modalities. The concatenated features are processed through a shared fully-connected classification head:

$$\hat{\mathbf{Y}} = \text{Softmax}(\mathbf{W}_{\text{FC}}\mathbf{U}), \tag{7}$$

where $\mathbf{W}_{\text{FC}} \in \mathbb{R}^{n_{\text{out}}\times n_{\text{class}}}$, $n_{\text{class}} = 4$ corresponds to the four collaboration satisfaction categories, and $\hat{\mathbf{Y}}$ represents the predicted class probabilities. We employ cross-entropy loss for training:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{4} y_{i,c}\log(p_{i,c}), \tag{8}$$

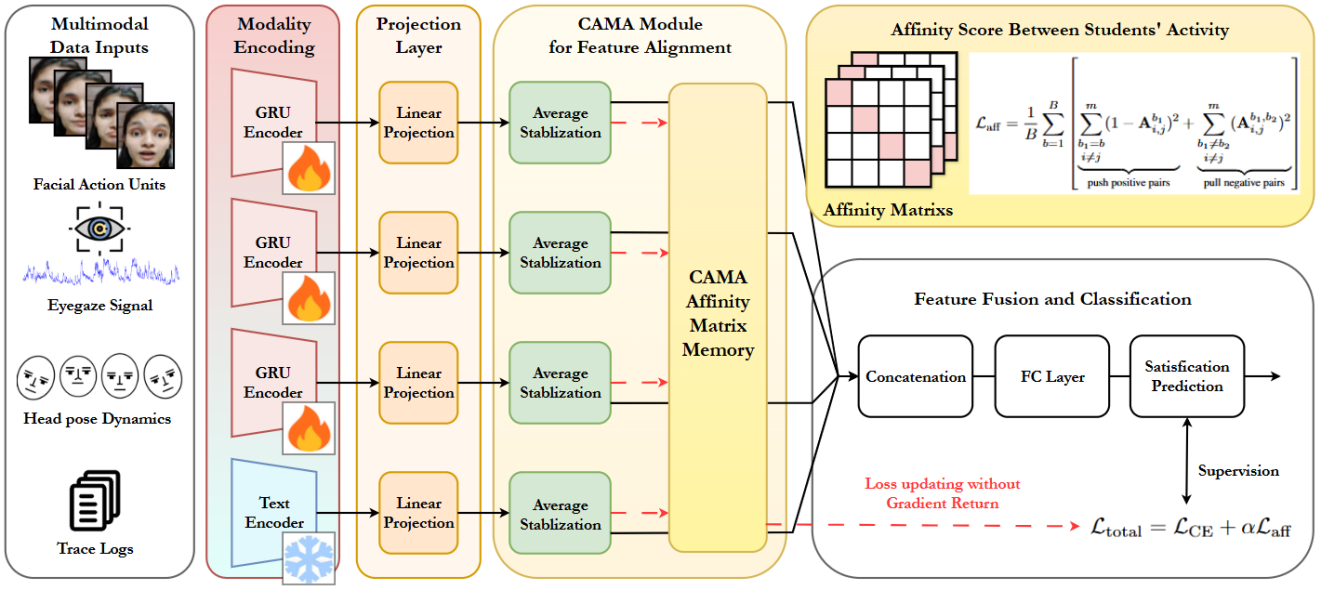


Figure 1. **Overview of the proposed AAMLA framework.** Four modality streams (facial action units, head pose, eye gaze, trace logs) are encoded by modality-specific encoders and projected into a unified $d = 128$ semantic space. The CAMA module explicitly models inter-modal relationships via affinity matrices and contrastive loss $\mathcal{L}_{\text{aff}}$, suppressing uninformative modalities. Aligned embeddings are classified by a FC head optimized with $\mathcal{L}_{\text{CE}}$, producing four-class collaboration satisfaction predictions.

where N is the number of training samples, $y_{i,c}$ is the ground-truth one-hot label, and $p_{i,c}$ is the predicted probability for class c.

3.3. Cross-modal Affinity-guided Alignment

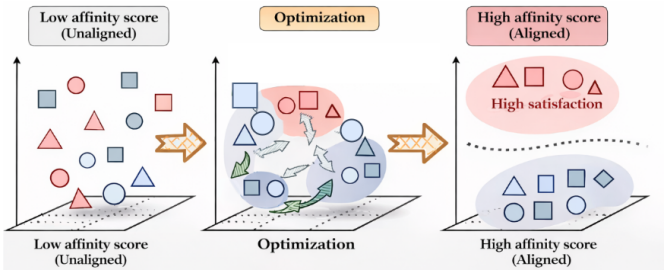


Figure 2. **Pipeline of the proposed CAMA strategy.** Different shapes denote different modalities; color denotes satisfaction class (red: high; blue: low). CAMA pulls same-class modality embeddings together and pushes apart different-class embeddings via affinity matrices, transforming scattered unaligned features (*left*) into compact, semantically coherent clusters (*right*) robust to uninformative modalities such as gaze.

Prior cross-attention fusion mechanisms [27] create critical vulnerabilities in multimodal student modeling: (1) over-concentration on dominant modalities, (2) lack of explicit inter-modal relationship modeling, and (3) cascading performance failures when dominant modalities carry noisy or inconsistent signals. In our setting, these vulnerabilities are particularly acute as modality informativeness varies significantly across student cohorts and interaction contexts [32].

To address these limitations, we propose the CAMA strategy, which explicitly models inter-modal relationships through affinity matrices. As illustrated in Fig. 2, CAMA enforces projected modality features into a compact and coherent shared space while ensuring semantic consistency even when individual modalities degrade (e.g., gaze occlusion, noisy AU detection, irregular trace sampling).

Given a batch of size B, batch-level modality embeddings $\mathbf{E}_{\text{AU}}, \mathbf{E}_{\text{pose}}, \mathbf{E}_{\text{gaze}}, \mathbf{E}_{\text{trace}} \in \mathbb{R}^{B\times d}$ are extracted using their respective encoders. To mitigate within-batch feature discrepancies caused by student behavioral variation, we compute batch-level averages across the degraded and non-degraded feature pairs for each modality:

$$\hat{\mathbf{E}}_{\text{AU}} = \tfrac{1}{2}(\mathbf{E}^{\text{orig}}_{\text{AU}} + \mathbf{E}^{\text{deg}}_{\text{AU}}), \quad \hat{\mathbf{E}}_{\text{pose}} = \tfrac{1}{2}(\mathbf{E}^{\text{orig}}_{\text{pose}} + \mathbf{E}^{\text{deg}}_{\text{pose}}),$$
$$\hat{\mathbf{E}}_{\text{gaze}} = \tfrac{1}{2}(\mathbf{E}^{\text{orig}}_{\text{gaze}} + \mathbf{E}^{\text{deg}}_{\text{gaze}}), \quad \hat{\mathbf{E}}_{\text{trace}} = \tfrac{1}{2}(\mathbf{E}^{\text{orig}}_{\text{trace}} + \mathbf{E}^{\text{deg}}_{\text{trace}}), \tag{9}$$

where superscripts orig and deg denote the original and degraded feature versions respectively, with degradation applied according to the modality degradation settings described in Sec. 4. The averaged embeddings $\hat{\mathbf{E}}_{\{\cdot\}} \in \mathbb{R}^{B\times d}$ provide stable representations irrespective of modality quality variations across students.

The stabilized embeddings are concatenated to form a joint multimodal representation:

$$\mathbf{U}^b = [\hat{\mathbf{E}}_{\text{AU}}, \hat{\mathbf{E}}_{\text{pose}}, \hat{\mathbf{E}}_{\text{gaze}}, \hat{\mathbf{E}}_{\text{trace}}] \in \mathbb{R}^{B\times 4d}, \tag{10}$$

enabling direct semantic comparison across all modalities. Unlike cross-attention fusion that suffers from cascading error accumulation under modality degradation, this concatenation strategy treats all modalities equally and eliminates intermediate computations that amplify alignment errors.

To explicitly quantify semantic correlations among modalities, we compute an affinity matrix:

$$\mathbf{A}^b = \sigma(\mathbf{U}^b(\mathbf{U}^b)^\top), \tag{11}$$

where σ is a normalization function bounding affinity scores within $[0, 1]$. This symmetric positive semi-definite matrix captures explicit semantic relationships among AU, pose, gaze, and trace features, ensuring robust modality-level alignment.

Guided by the affinity matrix, we introduce a contrastive learning strategy to enforce semantic alignment and enhance cross-modal consistency. Within each batch, positive pairs are formed from embeddings of different modalities belonging to the same student activity, promoting cross-modal coherence. Negative pairs are constructed from embeddings across different student activities, enhancing discriminative power. The affinity-driven contrastive loss is formulated as:

$$\mathcal{L}_{\text{aff}} = \frac{1}{B}\sum_{b=1}^{B}\left[\underbrace{\sum_{\substack{b_1=b\\ i\neq j}}^{m}(1-\mathbf{A}_{i,j}^{b_1})^2}_{\text{push positive pairs}} + \underbrace{\sum_{\substack{b_1\neq b_2\\ i\neq j}}^{m}(\mathbf{A}_{i,j}^{b_1,b_2})^2}_{\text{pull negative pairs}}\right], \tag{12}$$

where $\mathbf{A}_{i,j}^{b_1}$ represents the affinity score between modalities i and j for student activity b_1, and $\mathbf{A}_{i,j}^{b_1,b_2}$ denotes the cross-activity affinity score.

The proposed CAMA explicitly captures modality interactions and offers interpretable insights into cross-modal feature alignment across diverse student cohorts, even under modality degradation. Unlike prior implicit fusion approaches [27], CAMA employs explicit affinity matrices to model inter-modal relationships, mitigating the semantic inconsistencies that arise when gaze or other modalities carry inconsistent signals across students.

3.4. Training Objectives

The total loss integrates the cross-entropy classification loss $\mathcal{L}_{\text{CE}}$ with the affinity alignment loss $\mathcal{L}_{\text{aff}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{aff}}, \tag{13}$$

where α is empirically set to 0.25 via grid search. This formulation ensures that classification accuracy and cross-modal feature alignment are jointly optimized, leading to a more robust and generalizable collaboration satisfaction prediction framework.

4. Experiments Settings and Results

Dataset. We utilize multimodal data collected from 50 middle school students (6th–8th grade, ages 11–14) interacting with the EcoJourneys collaborative game-based learning environment [8], comprising 164 completed activities with paired video and trace log recordings. EcoJourneys is a problem-based collaborative learning environment in which small groups of students investigate the cause of an unknown illness affecting a fish population on a fictional Philippine island. As shown in Fig. 3, students are guided by the TIDE inquiry cycle (Talk, Investigate, Deduce, Explain): they converse with non-player characters (NPCs) to gather domain knowledge (Fig. 3b), explore the virtual island to collect evidence, and engage in collaborative reasoning via an in-game chat interface and a shared virtual whiteboard (Fig. 3a). Students progress through four activities (one tutorial and three quests), with an exit survey administered after each activity to capture their sentiments regarding the collaborative experience.

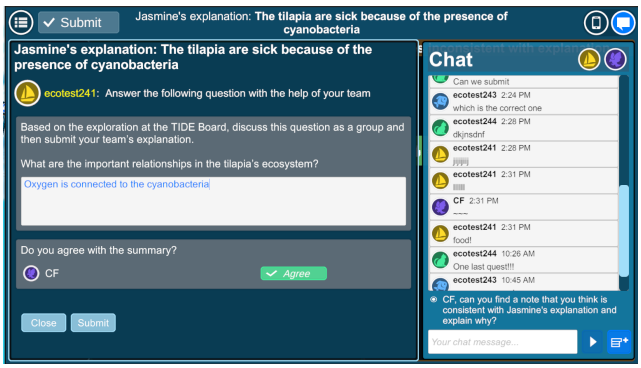
Video data was captured via front-facing laptop cameras during gameplay, while trace logs documented fine-grained in-game actions including NPC interactions, locations visited, evidence collected, and chat messages exchanged. Collaboration satisfaction labels were derived from the Likert-scale exit surveys, categorized into four classes based on group listening and idea-building responses. We follow a student-level 10-fold cross-validation protocol to ensure that all activities from the same student appear exclusively in either training or testing splits.

Feature Extraction and Synchronization. Video-based features are extracted using OpenFace 2.0 [4] from front-facing camera recordings, yielding three complementary facial feature streams: (1) 17 AU intensity values per frame ranging from 0 (absent) to 5 (high intensity); (2) 6 head pose features encoding 3D translation and rotation coordinates; and (3) 6 gaze direction features encoding 3D coordinates per eye. Trace log events are encoded as 768-dimensional embeddings using a BERT-based sentence encoder fine-tuned via unsupervised learning on in-game text [27], with a byte-pair encoding tokenizer to capture semantic nuances specific to in-game interactions.

Since facial features and trace log events are sampled at different rates, we follow [1] and adopt trace log events as the base sampling rate, synchronizing facial features by averaging values between successive trace events. Formally, denoting facial features as $\mathbf{F}(t)$ and trace log events as $\mathbf{S}(t_k)$, where t_k represents the slower trace sampling rate, synchronized facial features $\bar{\mathbf{F}}(t_k)$ are computed as:

$$\bar{\mathbf{F}}(t_k) = \frac{1}{|\mathcal{I}_k|}\sum_{t\in\mathcal{I}_k}\mathbf{F}(t), \tag{14}$$

where $\mathcal{I}_k$ denotes the interval $(t_{k-1}, t_k]$ and $|\mathcal{I}_k|$ is the count



(a) The TIDE Board interface, where students collaboratively construct and submit explanations alongside real-time group chat, generating rich trace log data.

(b) The 3D exploration view, where students navigate the virtual island, interact with NPCs such as local farmers, and coordinate actions through the in-game chat interface.

Figure 3. **Student communication while playing in the EcoJourneys collaborative learning environment [1].** Students work in small groups to investigate a fish illness on a virtual Philippine island, generating rich multimodal behavioral signals — including facial expressions, head pose, eye gaze, and in-game chat interactions — that our AAMLA framework leverages for collaboration satisfaction prediction. Unlike prior implicit fusion approaches, AAMLA explicitly aligns these heterogeneous modalities to suppress uninformative signals and produce robust cross-modal representations.

of facial feature frames within that interval. This yields a synchronized dataset wherein each trace event $\mathbf{S}(t_k)$ corresponds to a mean facial feature vector $\bar{\mathbf{F}}(t_k)$, forming the multimodal input sequences fed into our modality-specific encoders.

Modality Degradation Settings. To systematically evaluate robustness against modality degradation — a core motivation of our work — we simulate three degradation conditions following prior work on missing and corrupted modalities [23, 30, 45]:

- **Gaze dropout**: randomly zeroing gaze features at rates of 30%, 50%, and 70% of timesteps, simulating occlusion or tracking failure.
- **Gaussian perturbation**: adding noise $\mathcal{N}(0, \sigma)$ with $\sigma \in \{0.01, 0.05\}$ to AU and pose features, simulating sensor noise and compression artifacts.
- **Full modality absence**: completely removing one modality at inference, evaluating graceful degradation without retraining.

These settings reflect realistic challenges in educational deployments, where behavioral variation and hardware limitations cause modality quality to fluctuate across students and sessions.

Hyperparameter Settings and Implementation Details. Models are trained with a batch size of 16 using the Adam optimizer [14] with an initial learning rate of 1×10^{-3}, reduced on validation loss plateau. Early stopping with patience of 3 is applied for regularization, with a maximum of 500 epochs. The affinity alignment weight α is set to 0.25, selected via grid search over $\{0.1, 0.25, 0.5\}$. Dropout of 0.1 is applied to all GRU encoders. All modality-specific features are projected into a unified $d = 128$ dimensional space via learnable linear projection layers prior to affinity-driven alignment. Data synchronization follows the original protocol [8], averaging facial features between successive trace events. Evaluation metrics include macro F1-score and accuracy, consistent with the class-imbalanced nature of the satisfaction labels. All experiments are implemented in PyTorch and run on a single NVIDIA A100 GPU.

4.1. Unimodal Baseline Results

We first evaluate the predictive performance of each modality independently using GRU-based unimodal models, establishing baseline benchmarks consistent with prior work [1]. As shown in Table 1, all unimodal models achieve significant improvement over a majority classifier, demonstrating that each modality captures meaningful cues for collaboration satisfaction prediction. AU and pose features yield the strongest unimodal performance (F1 = 0.66, Acc = 0.66), while gaze and trace features perform comparably (F1 = 0.65, Acc = 0.65). Wilcoxon signed-rank tests across cross-validation folds reveal no significant performance differences among most unimodal models, with the exception of pose versus gaze ($p < .05$), suggesting that these two modalities capture complementary but asymmetric information about collaborative dynamics.

4.2. Multimodal Ablation Study

To understand the contribution of each component in our framework, we conduct a comprehensive ablation study comparing AAMLA against the prior cross-attention baseline [1] and intermediate model variants. Results are reported in Table 2.

Model A reproduces the prior cross-attention base-

Table 1. Comparison of unimodal baselines, cross-attention [1], and AAMLA. Mean ± std. over 10 runs of student-level 10-fold cross-validation.

Model	F1-Score	Accuracy
AU (Unimodal)	0.66 ±0.03	0.66 ±0.03
Pose (Unimodal)	0.66 ±0.04	0.66 ±0.04
Gaze (Unimodal)	0.65 ±0.05	0.65 ±0.05
Trace (Unimodal)	0.65 ±0.04	0.65 ±0.04
Cross-Attention [1]	0.72 ±0.03	0.72 ±0.03
AAMLA (Ours)	**0.79** ±0.02	**0.77** ±0.02

Table 2. Ablation study comparing model components. Proj.: modality projection layers. CAMA: Cross-modal Affinity-guided Modality Alignment. $\mathcal{L}_{\text{aff}}$: contrastive alignment loss.

Model	Proj.	CAMA	$\mathcal{L}_{\text{aff}}$	F1	Acc
A: Cross-Attention [1]	✗	✗	✗	0.72	0.72
B: + Projection	✓	✗	✗	0.71	0.72
C: + Projection + CAMA	✓	✓	✗	0.75	0.74
D: + Projection + CAMA + $\mathcal{L}_{\text{aff}}$ (Full)	✓	✓	✓	0.79	0.77

line [1], serving as the primary point of comparison. Model B isolates the effect of projection-based feature space unification, examining whether aligning heterogeneous modality dimensions into a common $d = 128$ space alone improves performance. Model C adds the CAMA affinity matrix fusion without contrastive supervision, assessing the contribution of explicit inter-modal relationship modeling. Model D represents the full AAMLA framework, incorporating all three components.

We additionally reproduce the modality combination ablation from [1] under our framework to examine whether CAMA changes which modality combinations are most effective, as reported in Table 3.

Table 3. Modality combination ablation under AAMLA framework, compared to prior cross-attention results [1].

Modality Combination	Cross-Attn [1]		AAMLA (Ours)	
	F1	Acc	F1	Acc
Full Multimodal	0.72	0.72	0.79	0.77
Trace + AU	0.70	0.70	0.73	0.74
Trace + Pose	0.69	0.69	0.72	0.74
Trace + Gaze	0.65	0.65	0.70	0.71
AU + Pose	0.68	0.67	0.71	0.72
AU + Gaze	0.65	0.65	0.70	0.71
Pose + Gaze	0.64	0.64	0.69	0.68
Trace + AU + Pose	0.70	0.70	0.74	0.76

4.3. Modality Robustness Evaluation

A critical advantage of AAMLA over the prior cross-attention approach [1] is its explicit mechanism for handling uninformative or degraded modalities. We systematically evaluate robustness under the three degradation conditions defined in Sec. 4, comparing AAMLA against the cross-attention baseline [1]. Results are summarized in Table 4.

Table 4. **Modality robustness evaluation** comparing Cross-Attention [1] and AAMLA (Ours) under three degradation conditions. Results reported as macro F1-score.

Modality	Degradation	Cross-Attn [1]	AAMLA (Ours)
Gaze	Dropout 30%	0.68	0.77
	Dropout 50%	0.61	0.75
	Dropout 70%	0.52	0.72
AU + Pose	$\mathcal{N}(0, 0.01)$	0.65	0.78
	$\mathcal{N}(0, 0.05)$	0.54	0.72
Full Absence	w/o AU	0.54	0.73
	w/o Pose	0.53	0.72
	w/o Gaze	0.52	0.68
	w/o Trace	0.56	0.70

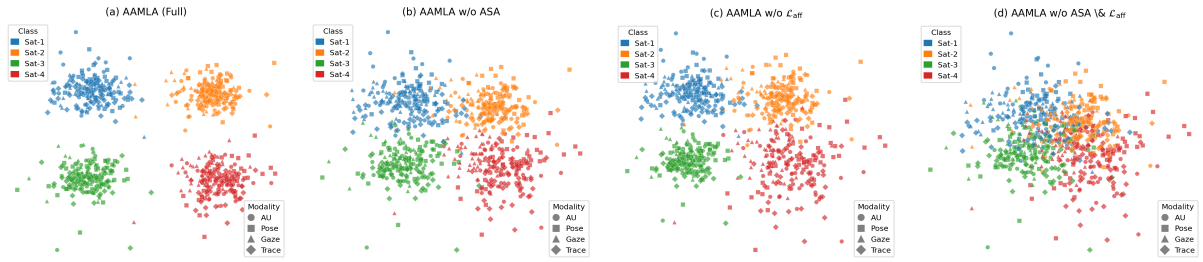
As demonstrated in the deepfake detection domain [22], explicit affinity-based alignment creates inherent redundancy among modalities such that when one modality degrades, learned cross-modal relationships allow the remaining modalities to compensate. We hypothesize that AAMLA exhibits analogous behavior in the educational domain, particularly for gaze dropout conditions where the prior cross-attention model [1] is known to be sensitive.

4.4. Analyses

t-SNE Feature Distribution. Fig. 4 presents t-SNE visualizations under four ablation settings. The full AAMLA model [Fig. 4(a)] produces well-separated clusters across the four satisfaction classes, with modality-specific embeddings remaining semantically cohesive. Removing CAMA [Fig. 4(b)] introduces visible cross-modality drift and inter-class overlap, reflecting weaker inter-modal consistency. Without $\mathcal{L}_{\text{aff}}$ [Fig. 4(c)], cluster boundaries become fragmented with uneven variance across cohorts. The fully unaligned model [Fig. 4(d)] yields the most scattered feature space, where class boundaries blur and inter-modality gaps widen substantially. Together, these results confirm that CAMA and $\mathcal{L}_{\text{aff}}$ jointly contribute to compact, discriminative representations.

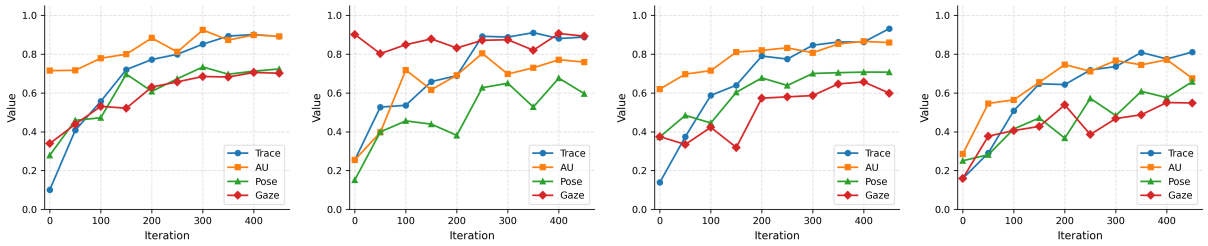
Affinity Score Evolution. Fig. 5 shows affinity score trajectories during training. Trace embeddings converge most stably, consistent with their semantic richness as a behavioral modality, while AU and pose exhibit more dynamic patterns reflecting their interaction dependencies. Gaze features show the lowest and most variable scores, corroborating prior observations of their inconsistent informativeness across student cohorts [1]. High-satisfaction activities reach stable alignment earlier than low-satisfaction ones, suggesting that positive collaborative interactions produce more consistent cross-modal patterns.

SHAP Analysis. As shown in Fig. 6, trace log features (`Chat_freq`, `NPC_Interact`, `Evidence_collect`)



(a) AAMLA (Full) (b) AAMLA w/o CAMA (c) AAMLA w/o $\mathcal{L}_{\text{aff}}$ (d) AAMLA w/o CAMA & $\mathcal{L}_{\text{aff}}$

Figure 4. **t-SNE visualizations of multimodal feature distributions under different ablation settings.** Color denotes satisfaction class (Sat-1 to Sat-4); marker shape denotes modality (AU, Pose, Gaze, Trace). (a) The full AAMLA model produces tightly clustered, semantically aligned features with clear inter-class separation. (b) Removing CAMA causes cross-modality drift and partial overlap between satisfaction classes. (c) Removing $\mathcal{L}_{\text{aff}}$ produces fragmented decision boundaries with increased cohort-level variance. (d) Removing both components yields the most scattered feature space, with blurred inter-class boundaries and widened inter-modality semantic gaps.



(a) High-satisfaction affinity scores (b) Low-satisfaction affinity scores (c) Degraded high-satisfaction affinity scores (d) Degraded low-satisfaction affinity scores

Figure 5. **Affinity scores evolution among AU, pose, gaze, and trace modalities across {high-satisfaction, low-satisfaction} and {original, degraded} conditions during training.** High-satisfaction activities achieve stable convergence earlier, reflecting more consistent cross-modal alignment, while degraded conditions exhibit higher variance, particularly for gaze features, motivating the explicit alignment enforced by CAMA.

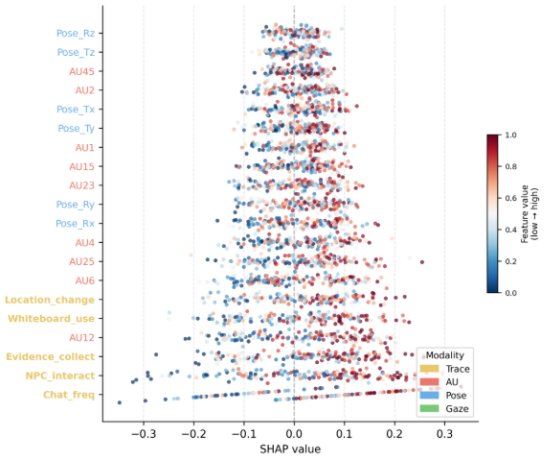


Figure 6. **SHAP beeswarm plot of feature contributions**. Color denotes normalized feature value. Trace features rank highest, gaze features are absent from the top-20, corroborating CAMA's adaptive suppression of uninformative modalities.

consistently rank highest, confirming their dominance in satisfaction prediction. AU and pose features show moderate contributions, while gaze features are absent from the top-20 — directly validating CAMA's adaptive suppression of uninformative modalities.

5. Conclusion

In this paper, we presented AAMLA, an Affinity-Aligned Multimodal Learning Analytics framework that addresses modality degradation in collaborative game-based learning through Cross-modal Affinity-guided Modality Alignment (CAMA). By explicitly modeling inter-modal relationships via affinity matrices and contrastive learning, CAMA adaptively suppresses uninformative modalities such as gaze without discarding them, producing robust representations that generalize across diverse student cohorts. Experiments on EcoJourneys confirm consistent improvements over unimodal baselines and the prior cross-attention approach under both standard and degraded conditions, with SHAP and t-SNE analyses providing interpretable insights into the behavioral signals most predictive of collaboration satisfaction. More broadly, this work demonstrates that explicit multimodal alignment techniques transfer effectively to educational settings where modality reliability is inherently variable.

References

[1] Halim Acosta, Seung Lee, Bradford Mott, Haesol Bae, Krista Glazewski, Cindy Hmelo-Silver, and James C. Lester. Multimodal learning analytics for predicting student collaboration satisfaction in collaborative game-based learning. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 224–235, Atlanta, Georgia, USA, 2024. International Educational Data Mining Society. 2, 5, 6, 7

[2] Waleed Mugahed Al-Rahmi and Mohd Shahizan Othman. Evaluating student's satisfaction of using social media through collaborative learning in higher education. *International Journal of Advances in Engineering & Technology*, 6 (4):15–41, 2013. 1, 2

[3] Youngkyun Baek and Ahmed Touati. Comparing collaborative and cooperative gameplay for academic and gaming achievements. *Journal of Educational Computing Research*, 57(8):2110–2140, 2020. 2

[4] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 IEEE International Conference on Automatic Face and Gesture Recognition*, pages 59–66. IEEE, 2018. 2, 3, 5

[5] Nuno Borges, Leif Lindblom, Benjamin Clarke, Anna Gander, and Rodney Lowe. Classifying confusion: Autodetection of communicative misunderstandings using facial action units. In *2019 Affective Computing and Intelligent Interaction Workshops and Demos*, pages 401–406, 2019. 3

[6] Matthew Bradford, Imene Khebour, Nathan Blanchard, and Nirmalya Krishnaswamy. Automatic detection of collaborative states in small groups using multimodal features. In *Proceedings of the 24th International Conference on Artificial Intelligence in Education*, pages 767–773, 2023. 2

[7] Mengxue Cai and Clayton D. Epp. Modeling cognitive load and affect to support adaptive online learning. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 799–804, 2022. 2

[8] Dustin Carpenter, Andrew Emerson, Bradford W. Mott, Abeer Saleh, Krista D. Glazewski, Cindy E. Hmelo-Silver, and James C. Lester. Detecting off-task behavior from student dialogue in game-based collaborative learning. In *Artificial Intelligence in Education: 21st International Conference*, pages 55–66. Springer, 2020. 2, 3, 5, 6

[9] Pankaj Chejara, Luis P. Prieto, María J. Rodríguez-Triana, Ángel Ruiz-Calleja, Riin Kasepalu, Ioanna-Angeliki Chounta, and Barbara Schneider. Exploring indicators for collaboration quality and its dimensions in classroom settings using multimodal learning analytics. In *European Conference on Technology Enhanced Learning*, pages 60–74. Springer, 2023. 3

[10] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. In *Advances in Neural Information Processing Systems*, 2021. 1, 2

[11] Xiaoqing Chen, Di Zou, Haoran Xie, Gong Cheng, and Fanyu Su. A bibliometric analysis of game-based collaborative learning between 2000 and 2019. *International Journal of Mobile Learning and Organisation*, 16(1):20–51, 2022. 2

[12] Imen Daoudi, Emmanuel Tranvouez, Rihab Chebil, Bernard Espinasse, and Wajdi L. Chaari. An EDM-based multimodal method for assessing learners' affective states in collaborative crisis management serious games. In *Proceedings of the 13th International Conference on Educational Data Mining*, 2020. 2

[13] Yael Dich, Jennifer M. Reilly, and Barbara Schneider. Using physiological synchrony as an indicator of collaboration quality, task performance and learning. In *Artificial Intelligence in Education: 19th International Conference, AIED 2018*, pages 98–110. Springer, 2018. 1

[14] P. Kingma Diederik and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 6

[15] Iván Fonseca, Manuel Caviedes, Juan Chantré, and Jaime Bernate. Gamification and game-based learning as cooperative learning tools: A systematic review. *International Journal of Emerging Technologies in Learning (iJET)*, 18(21): 4–23, 2023. 1, 2

[16] Jingsheng Gao, Jiacheng Ruan, Suncheng Xiang, Zefang Yu, Ke Ji, Mingye Xie, Ting Liu, and Yuzhuo Fu. Lamm: Label alignment for multi-modal prompt learning. *arXiv preprint arXiv:2312.08212*, 2023. 2

[17] Amy E. Griffith, Grace A. Katuka, Joseph B. Wiggins, Kristy E. Boyer, Jason Freeman, Brian Magerko, and Taylor McKlin. Investigating the relationship between dialogue states and partner satisfaction during co-creative learning tasks. *International Journal of Artificial Intelligence in Education*, 33(3):543–582, 2023. 2, 3

[18] Zhongyang Guo and Reza Barmaki. Deep neural networks for collaborative learning analytics: Evaluating team collaborations using student gaze point prediction. *Australasian Journal of Educational Technology*, 36(6):53–71, 2020. 1, 2, 3

[19] Celeste B. Harris, Penny Van Bergen, Samantha A. Harris, Natalie McIlwain, and Aline Arguel. Here's looking at you: eye gaze and collaborative recall. *Psychological Research*, 86:769–779, 2022. 1, 2, 3

[20] Kenan Hava, Tolga Guyer, and Hasan Cakir. Gifted students' learning experiences in systematic game development process in after-school activities. *Educational Technology Research and Development*, 68:1439–1459, 2020. 1

[21] Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, and Wen Zhang. Structure-clip: Towards scene graph knowledge to enhance multi-modal structured representations. In *Proc. AAAI Conference on Artificial Intelligence*, 2024. 2

[22] Ching-Yi Lai, Chih-Yu Jian, Pei-Cheng Chuang, Chia-Ming Lee, Chih-Chung Hsu, Chiou-Ting Hsu, and Chia-Wen Lin. Umcl: Unimodal-generated multimodal contrastive learning for cross-compression-rate deepfake detection. *International Journal of Computer Vision*, 134:40, 2026. 2, 7

[23] Binh M. Le and Simon S. Woo. Quality-agnostic deepfake detection with intra-model collaborative learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22321–22332, 2023. 2, 6

[24] Jiyoung Lee, Soo-Whan Chung, Sunok Kim, Hong-Goo Kang, and Kwanghoon Sohn. Looking into your speech:

Learning cross-modal affinity for audio-visual speech separation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[25] Soo Jeoung Lee, Siva Srinivasan, Tracy Trail, David Lewis, and Suzanne Lopez. Examining the relationship among student perception of support, course satisfaction, and learning outcomes in online learning. *The Internet and Higher Education*, 14(3):158–163, 2011. 1, 2

[26] Jie Li, Yu Lin, Ming Sun, and Rustam Shadiev. Socially shared regulation of learning in game-based collaborative learning environments promotes algorithmic thinking, learning participation and positive learning attitudes. *Interactive Learning Environments*, 31(3):1715–1726, 2023. 2

[27] Peizhong Li, Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, Rahul Jain, Varun Manjunatha, and Hui-juan Liu. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660, 2021. 1, 2, 3, 4, 5

[28] Hsin-Yu Liang, Tsung-Yen Hsu, Gwo-Jen Hwang, Shih-Chun Chang, and Hsiao-Chen Chu. A mandatory contribution-based collaborative gaming approach to enhancing students' collaborative learning outcomes in science museums. *Interactive Learning Environments*, 31(5):2692–2706, 2023. 2

[29] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, 2022. 2

[30] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proc. AAAI Conference on Artificial Intelligence*, pages 2302–2310, 2021. 1, 2, 6

[31] Yuxin Ma, Mehmet Celepkolu, and Kristy Elizabeth Boyer. Detecting impasse during collaborative problem solving with multimodal learning analytics. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 45–55, 2022. 2

[32] Yuxin Ma, Grace A. Katuka, Mehmet Celepkolu, and Kristy Elizabeth Boyer. Investigating multimodal predictors of peer satisfaction for collaborative coding in middle school. In *Proceedings of the 15th International Conference on Educational Data Mining*. International Educational Data Mining Society, 2022. 2, 3, 4

[33] Bradley McDaniel, Sidney D'Mello, Brent King, Patrick Chipman, Kristopher Tapp, and Arthur Graesser. Facial features for affective state detection in learning environments. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2007. 3

[34] Jennifer K. Olsen, Kshitij Sharma, Nikol Rummel, and Vincent Aleven. Temporal analysis of multimodal data to predict collaborative learning outcomes. *British Journal of Educational Technology*, 51(5):1527–1547, 2020. 2, 3

[35] Satyapriya Praharaj, Maren Scheffel, Martin Schmitz, Marcus Specht, and Hendrik Drachsler. Towards collaborative convergence: quantifying collaboration quality with automated co-located collaboration analytics. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 358–369, 2022. 2

[36] Roberto U. Puga. Game-based learning: a tool that enhances the collaborative work. In *European Conference on Games Based Learning*, pages 570–577, 2022. 2

[37] Barbara Schneider and Roy Pea. Toward collaboration sensing. *International Journal of Computer-Supported Collaborative Learning*, 9:371–395, 2014. 1

[38] Kshitij Sharma, Ioannis Leftheriotis, and Michail Giannakos. Utilizing interactive surfaces to enhance learning, collaboration and engagement: Insights from learners' gaze and speech. *Sensors*, 20(7):1964, 2020. 3

[39] Hyo-Jeong So and Thomas A. Brush. Student perceptions of collaborative learning, social presence and satisfaction in a blended learning environment: Relationships and critical factors. *Computers & Education*, 51(1):318–336, 2008. 1, 2

[40] Emily L. Starr, Jennifer M. Reilly, and Barbara Schneider. Toward using multi-modal learning analytics to support and measure collaboration in co-located dyads. In *ICLS 2018: 13th International Conference of the Learning Sciences*, pages 448–455. International Society of the Learning Sciences, 2018. 2

[41] Andrew E. Stewart, Zachary Keirn, and Sidney K. D'Mello. Multimodal modeling of collaborative problem-solving facets in triads. *User Modeling and User-Adapted Interaction*, 31(4):713–751, 2021. 2

[42] Özgür Sümer, Paul Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*, 14(2):1012–1027, 2021. 3

[43] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (11), 2008. 2

[44] Chao Wang and Lijuan Huang. A systematic review of serious games for collaborative learning: Theoretical framework, game mechanic and efficiency assessment. *International Journal of Emerging Technologies in Learning*, 16(6): 88–105, 2021. 2

[45] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023. 1, 2, 6

[46] Zehan Wang, Yang Zhao, Xize Cheng, Haifeng Huang, Jiageng Liu, Aoxiong Yin, Li Tang, Linjun Li, Yongqi Wang, Ziang Zhang, and Zhou Zhao. Connecting multi-modal contrastive representations. In *Advances in Neural Information Processing Systems*, 2023. 1, 2

[47] Yi Xin, Junlong Du, Qiang Wang, Ke Yan, and Shouhong Ding. Mmap: multi-modal alignment prompt for cross-domain multi-task learning. In *Proc. AAAI Conference on Artificial Intelligence*, 2024. 2

[48] Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time adaptation against multi-modal reliability bias. In *International Conference on Learning Representations*, 2024. 2

[49] Zhonggen Yu, Ming Gao, and Lili Wang. The effect of educational games on learning outcomes, student motivation, engagement and satisfaction. *Journal of Educational Computing Research*, 59(3):522–546, 2021. 2

[50] Abdulazeez Abubakar Yunusa and Ibraheem Nasirudeen Umar. A scoping review of critical predictive factors (CPFs) of satisfaction and perceived learning outcomes in e-learning environments. *Education and Information Technologies*, 26: 1223–1270, 2021. 2

[51] Johanna Zambrano, Paul A. Kirschner, and Femke Kirschner. How cognitive load theory can be applied to collaborative learning. In *Advances in Cognitive Load Theory: Rethinking Teaching*, pages 30–40. 2019. 1

[52] Chang Zhu. Student satisfaction, performance, and knowledge construction in online collaborative learning. *Journal of Educational Technology & Society*, 15(1):127–136, 2012. 1