Multi-Agent Reflection Loop: A Multimodal Framework for Cross-Cultural Narrative Adaptation

Anonymous ACL submission

Abstract

Cross-cultural narrative understanding demands language models to not only excel at text generation, but also perceive users' implicit cultural cognitive states and dynamically align with them. However, existing interactive narrative systems predominantly rely on static preset scripts or monolithic large language models (LLMs). These approaches fail to resolve semantic conflicts arising from multimodal user inputs (text, behavior, emotion), leading to feedback lacking cultural adaptability and even the generation of contextual illusions. To address these limitations, this paper proposes a Multimodal Culture-Aware Multi-Agent System (MC-MAS), which achieves high-precision cross-cultural narrative adaptation through a collaborative agent mechanism. Specifically, we design three specialized functional agents (behavioral, linguistic, cultural) to process heterogeneous user signals, and introduce a core coordinator agent. The coordinator employs a novel "reflection-reconstruction" loop mechanism, which can automatically detect cross-modal consistency conflicts and iteratively optimize narrative generation strategies. We validate the MC-MAS framework in a narrative scenario rooted in the exile literature of the Weimar Republic. Experimental results demonstrate that, compared with static models and single LLMs, our method significantly enhances the accuracy of cultural context alignment while preserving narrative coherence, and effectively alleviates users' cognitive load.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in role-playing and interactive storytelling, enabling the construction of complex fictional worlds guided by user instructions (Bartle, 2004; Dietz et al., 2023). These advances have made LLMs promising engines for interactive narrative systems and serious games, particularly in culturally situated learning and heritage-related contexts (Mortara et al., 2014; Titus and Ng'ambi, 2023). However, when such narratives are situated in cross-cultural contexts, the challenge extends beyond linguistic fluency or plot coherence. Effective interaction requires the system to accurately interpret users' culturally grounded cognitive states and dynamically adapt narrative strategies to achieve cultural alignment, an issue widely recognized in cross-cultural game-based learning research (Jossan et al., 2021; O'Hagan, 2009).

Crucially, cultural misalignment does not always manifest as explicit misunderstanding in textual input(Shen et al., 2025; Li and Li, 2025). Instead, it often emerges through inconsistencies across multiple modalities(Pescuma et al., 2023; Yang et al., 2025; Wafa et al., 2025). A user's text may appear logically coherent, while their in-game behavior, interaction patterns, or facial expressions(Zhang et al., 2024; Fan et al., 2023; Cloude et al., 2022) reveal confusion toward culturally specific metaphors, such as religious allusions or historical analogies. Such phenomena are consistent with findings in multimodal cultural meaning-making and perception research, where cultural understanding is distributed across linguistic, behavioral, and visual channels rather than confined to text alone (Cabezas-García and Reimerink, 2022; Xie, 2025). These discrepancies reflect deeper cognitive conflicts shaped by cultural background rather than surface-level language errors. As a result, cross-cultural narrative adaptation inherently becomes a multimodal inference problem, aligning with recent work on emotion-aware and multimodal interpretation in interactive systems (Vistorte et al., 2024).

Existing interactive narrative systems face several limitations in cross-cultural contexts:

- **Limited multimodal understanding:** Most systems rely primarily on textual input, ignoring behavioral and affective signals.

- **Inability to reconcile conflicting cues:** When signals from different modalities contradict each other, monolithic models cannot resolve the conflicts effectively.

- **Insufficient evaluation of user understanding:** Current approaches focus on narrative fluency rather than whether the system correctly interprets users' cognitive and cultural states.

To address these issues, we propose the Multimodal Culture-Aware Multi-Agent System (MCMAS), which provides the following contributions:

- **Multimodal multi-agent framework:** MCMAS integrates specialized expert agents to analyze behavioral, linguistic, and cultural signals, overcoming the limitation of text-only models.

- **Coordinator with reflection mechanism:** A central coordinator reconciles conflicting outputs from different agents, enabling robust interpretation of inconsistent cues.

- **Human-level evaluation:** We validate MCMAS on a human-annotated benchmark, showing near-human performance in user profiling and cross-cultural intent recognition.

2 Related Work

Multi-Agent Collaboration and Simulation Environments Current education and professional training are shifting from single-agent to multi-agent collaborative environments, leveraging Generative AI to simulate complex interaction dynamics. SimClass (Zhang et al., 2025) constructs a comprehensive classroom environment with diverse peer roles, utilizing the Flanders Interaction Analysis System (FIAS) to reproduce authentic teaching interactions. In the medical field, MEDCO (Wei et al., 2024) and *Agent Hospital* (Li et al., 2024b) utilize expert-patient triads and the "Simulacrum-based Evolutionary Agent Learning" (SEAL) paradigm, respectively, to facilitate the evolution of interdisciplinary reasoning and diagnostic capabilities. This collaborative model has also been applied to the cultural domain; TRANSAGENTS (Wu et al., 2025) produces literary translations that emphasize cultural resonance through a multi-agent virtual company workflow (including editors and localization experts) and proposes the Monolingual Human Preference (MHP) metric, which transcends literal matching.

Multimodal Analytics and Cultural Semantic Alignment To achieve deeper cognitive and semantic understanding, research attention has turned to the deep fusion of heterogeneous data and the calibration of model values. In learning analytics, MSC-Trans (Xie et al., 2025) and Yan et al. (Yan et al., 2025) integrate data streams, such as facial expressions, postures, and logs, to resolve data asynchrony issues and accurately detect student engagement, while Yan et al. (Yan et al., 2025) further demonstrate the synergy of Generative AI in imputing missing longitudinal data. Addressing cultural consistency in language models, CultureLLM (Li et al., 2024a) effectively augments data using the World Values Survey. Meanwhile, MENAValues (Zahraei and Asgari, 2025) reveals value inconsistencies and "log-odds leakage" challenges during "cross-lingual value transfer" through benchmarking, highlighting the critical importance of capturing profound cultural nuances.

3 Method

This section presents a multimodal analysis framework based on a Multi-Expert Committee. Culture, Behavior, and Language experts analyze students' expressions, actions, and text in parallel, while a Referee Module fuses their outputs and resolves conflicts, enabling accurate modeling of cultural profiles and gaming personas for adaptive teaching.

3.1 Data Collection

Gardner's Theory of Multiple Intelligences argues that traditional standardized testing, largely based on written assessments, offers a reductive view of cognition and overlooks the diversity of student potential. Although seven intelligence dimensions are defined, practical system constraints require a more focused implementation. Therefore, we collect data from three primary modalities—facial expressions, linguistic articulations, and operational behaviors—to construct students' Cultural Profiles and Bartle gaming personas.

Data collection is organized into three phases corresponding to three curriculum sessions. In each phase, players complete four tasks designed according to the Bartle Taxonomy to elicit behaviors of Achievers, Explorers, Socializers, and Killers. Data

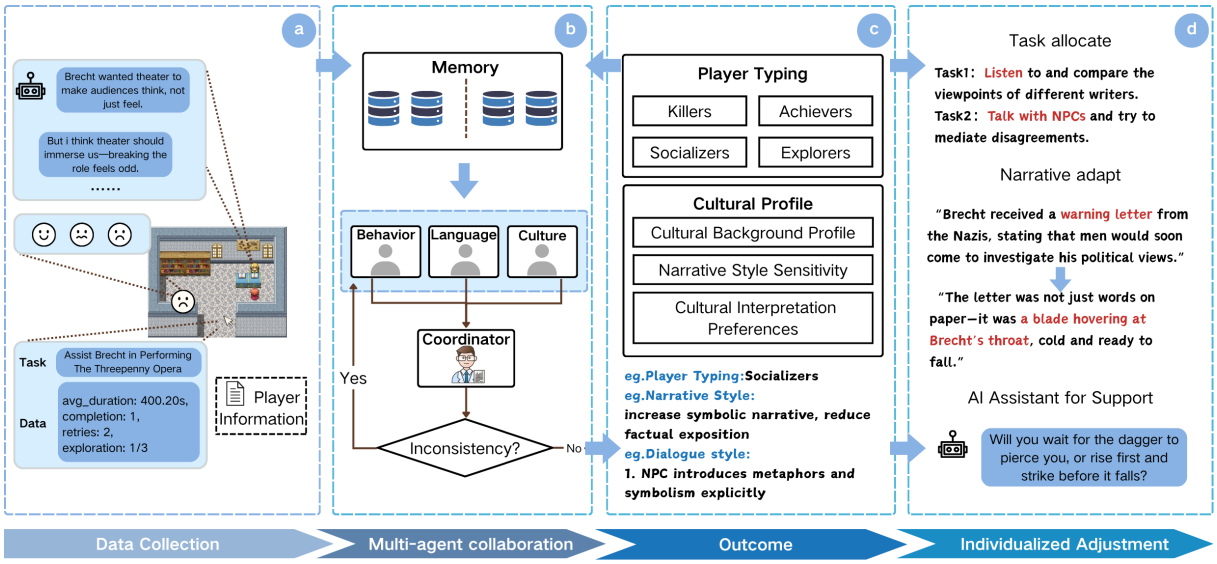


Figure 1: System workflow of the culturally adaptive game-based learning framework. (a) Data collection from player interactions, behavior logs, dialogues, and emotional cues; (b) multi-agent collaboration (behavior, language, culture agents with a coordinator for inconsistency resolution); (c) outcome generation including player typing and cultural profile; and (d) individualized adjustment through task allocation, narrative adaptation, and AI assistant support.

are analyzed sequentially, with outputs from each phase informing modeling in the next.

To capture facial expressions while preserving privacy, we adopt a discrete sequence sampling strategy. Instead of storing raw video, the system records emotion labels (e.g., Happy, Confused, Sad) at 5-second intervals, ensuring that no biometric facial data are retained. In parallel, we log fine-grained operational information, including task timestamps, failure counts, scene indices, AI dialogue history, and textual submissions. The collected data fields are detailed in Table 1.

Through this approach, we facilitate the unified temporal alignment of heterogeneous multimodal data streams, thereby laying a solid foundation for the subsequent analysis conducted by the Specialist Agents.

3.2 Workflow of Multi-Agent Reflective Cycles

Cultural adaptation in this system is implemented through a multi-agent collaborative mechanism. Learners' multimodal in-game inputs (Figure 1-a) are processed by specialized agents, and their outputs are integrated and reconciled by a central master agent. This expert-like deliberative design improves interpretability and consistency while enabling structured reflection and iterative optimization. The overall workflow is shown in Figure 1-b.

Behavior Analysis Agent The Behavior Agent (A_{beh}) specializes in transforming discrete game logs into high-level representations of player motivational profiles. We formally define the player's behavioral trajectory within a time window t as $\mathcal{B}_t = \{r_c, t_m, e_{ratio}, i_{freq}\}$.To bridge low-level metrics with psychological traits, A_{beh} functions as a probabilistic reasoner conditioned on behavioral psychology principles. Instead of simple rule-based mapping, the agent utilizes a Chain-of-Thought (CoT) reasoning process defined by our system instructions. It first interprets the raw metrics in the context of game mechanics (e.g., distinguishing "exploration" from "getting lost") and then maps these interpretations to the four Bartle player types (Killer, Achiever, Socializer, Explorer). The agent maximizes the conditional probability $P(\mathbf{P}_{bartle}|\mathcal{B}_t, \mathcal{K}_{psych})$, where $\mathcal{K}_{psych}$ represents the injected domain knowledge regarding player typologies. The output is a normalized distribution vector $\mathbf{P}_{bartle}$, quantifying the player's dominant and auxiliary motivations.

Language Analysis Agent The agent is tasked with analyzing the player's dialogue content and facial expressions during gameplay. By extracting key features that reflect the player's psychological state, it provides a basis for the dynamic adjustment of narrative content. We quantify the player's expressive patterns through three dimen-

Table 1: Selected Interaction Log: Brecht's Study Scenario

Time	Facial Expr.	Scene	User / AI Interaction	Input Buffer
0–5s	Neutral	Brecht's living room	None	None
...				
60–65s	Anxious	Brecht's study	User: "Is it... immersive experience?"	"Immersion"
70–75s	Happy	Brecht's study	User: "Ah! The alienation effect? Breaking the fourth wall?"	"Alienation"
...				
115–120s	Happy	Brecht's study	AI: "Task completed. Your understanding is correct."	None

sions: Metaphorical Density, Syntactic Complexity, and Linguistic Profile (High/Low Context).

To implement this, the agent aligns the player's text sequence (S_{text}) with facial expression labels (S_{face}). The feature extraction process is governed by a structured analysis protocol inspired by sociolinguistic frameworks. Specifically, the model is instructed to:Detect semantic alignment between player utterances and in-game lore to evaluate immersion;Parse syntactic structures to infer cognitive load (e.g., identifying fragmentation as a sign of confusion);Classify communication styles based on information density. Formally, the extraction can be denoted as $\mathcal{P}_{ling} = \text{Model}(S_{text}, S_{face}|I_{instruct})$, where $I_{instruct}$ encapsulates the definitions of the aforementioned dimensions. The resulting vector $\mathcal{P}_{ling}$ serves as a real-time "probe," distinguishing between states of *high immersion* and *cognitive overload*.

Cultural Context Analysis Agent The Cultural Agent, denoted as A_{cult}, functions as a bridge between the player's identity background and their in-game actions, effectively acting as a cross-cultural communication and historical consultant. It integrates the player's static profile $I_{profile}$ (e.g., nationality, native language) with a dynamic behavior log $\mathcal{B} = \{r_c, t_m, e_{ratio}, i_{freq}\}$ to construct a dynamic cultural model.

The core mechanism of A_{cult} involves Cultural Incongruity Detection. The agent integrates the static profile $I_{profile}$ with dynamic behavioral data to identify friction points. The reasoning logic follows a cause-effect derivation: it first scans for specific triggers (e.g., religious metaphors, historical allusions), checks the player's reaction (e.g., hesitation, negative sentiment), and then evaluates whether the disconnect stems from a cultural knowledge gap.

The output mechanism is two-fold:

- Risk Quantification ($O_{\text{RiskAlert}}$): A discrete classification (High/Medium/Low) predicting the likelihood of misunderstanding.
- Intervention Strategy ($O_{\text{BridgingSuggestion}}$): A generative task where the agent proposes specific remediation strategies, such as adding explanatory asides or simplifying cultural references, tailored to the specific type of detected incongruity.

Coordinator Agent: The Committee Chair Multimodal modeling inevitably engenders conflicts among agents. Sole reliance on initial inferences may lead to strategic distortion. To address this, we introduce a Reflection Round Mechanism that resolves multimodal conflicts through iterative feedback and model reconstruction. To address this, we introduce a Reflection-Arbitration Mechanism that strictly follows an "Identify-Remodel-Fuse-Update" loop:

Conflict Identification: The Coordinator aggregates outputs from sub-agents and calculates semantic discrepancies. A reflection round is triggered solely when conflicts exceed a pre-defined severity threshold.

Context-Aware Prompt Construction: Instead of generic re-prompting, the Coordinator dynamically synthesizes a meta-prompt. This prompt includes the conflicting conclusions, original evidence, and specific directives for re-evaluation, effectively contextualizing the reflection task.

Iterative Remodeling: Specialized agents reprocess their inputs under the guidance of the meta-prompt. For instance, the Behavioral Agent may incorporate cultural context provided by the Coordinator to re-interpret a "retry" action not as failure, but as curiosity.

Revision Fusion & Decision: The Coordinator performs a final weighted aggregation of the revised outputs to generate the adaptive decision D_{final}.

This mechanism ensures that the final decision is not merely a statistical average but a result of deliberative consensus, balancing interpretability with robustness against single-modal biases. The detailed prompt specifications for each agent are provided in Appendix B.

3.3 Task and Expression Intelligent Modification

To ensure a tiered learning objective structure, we referenced Bloom's Taxonomy(Blyth et al., 1966; Ormell, 1974) to divide each character's learning tasks into three stages:

Stage One (Remember and Understand): Players complete initial tasks to familiarize themselves with the game's fundamentals. This stage features four task types aligned with Bartle Player Types(Bartle, 2004)(Achiever, Explorer, Socializer, Killer) (Table 2) to cater to diverse learner preferences. During this process, the system collects multimodal player data, including text inputs, facial expressions, and operational behaviors.

Stages Two and Three (Apply and Analyze)(Evaluate and Create): The system employs a hierarchical dynamic adaptation mechanism (Figure 2). Vertically, data collected in prior stages feeds into subsequent stages to holistically adjust task allocation, narrative style, and difficulty design (inter-stage adaptation). Horizontally, the AI assistant receives real-time analysis results from the current stage, enabling instant adjustments to dialogue strategies and task prompts (intra-stage adaptation). For instance, the system increases annotation points and background extension tasks for exploration-oriented players, while introducing culturally contextualized explanations in narratives for players from diverse backgrounds.

This phased and personalized design not only deepens learners' comprehension of literary knowledge but also enhances learning adaptability across cultural contexts through dynamic regulation. Figure 1 illustrates the dynamic adaptation of in-game content. Specifically, Figure 1-a demonstrates the personalized task allocation tailored to player characteristics; Figure 1-b presents the script modifications informed by player profiles; and Figure 1-c highlights the individualized support offered by the AI system.

4 Experiment Setup

The following experimental setup is designed to assess both the inferential accuracy of the proposed multimodal and multi-agent reflective approach and its effects on players' cognitive load in the adaptive literary learning system.

4.1 Accuracy Evaluation

To validate the effectiveness of the Multi-Agent Reflection-Reconstruction Loop within the MC-MAS framework, specifically in the tasks of User Profiling and Cultural Intent Recognition, we designed a comparative experiment benchmarked against human expert annotations. The data collection protocols were reviewed and approved by the Institutional Review Board (IRB) of the authors' institution.

4.1.1 Dataset and Human Benchmark

Data Collection. Experimental data was derived from authentic gameplay logs. We recruited a total of 20 volunteer participants, consisting of undergraduate students majoring in German literature and enthusiasts with foundational proficiency in German (10 males, 10 females). From the raw dataset, we screened and selected 11 high-quality, complete interaction records (5 males, 6 females) to form the core test set. These records cover all three pedagogical phases of the game and contain unedited *Action Logs*, *Dialogue Text*, and *Facial Expression Sequences*. This multimodal dataset authentically reflects learners' cognitive states and operational habits when confronting cross-cultural narratives, constituting the foundational material for system evaluation.All interaction logs and behavioral data were anonymized prior to analysis and do not include any information that could be used to identify individual participants.

Ground Truth Annotation. To establish a reliable evaluation benchmark (Ground Truth), we assembled an "Expert Annotation Committee" comprising 11 domain experts: one professor of German literature and ten postgraduate students in the same field. We provided the committee with the

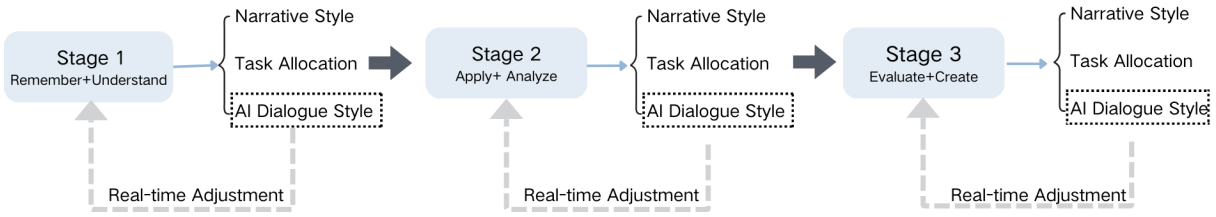


Figure 2: Cultural Adaptation for Segmented Operation. We divided the learning content of a writer into three stages according to Bloom's taxonomy of educational objectives: Remember and Understand, Apply and Analyze, Evaluate and Create. The analysis of Narrative Style, Task Allocation, and AI Dialogue Style in current stage will influence the next stage, while AI Dialogue Style also exerts effects on the current stage.

Table 2: Player types and staged tasks used in our game-based study. Tasks are organized from Task 1 to Task 5.

Player Type	**Task 1**	**Task 2**	**Task 3**	**Task 4**	**Task 5**
Achiever	Level–challenge task	Achievement–unlock task	Time–limited objective	Statistics–oriented task	Completionist collection
Explorer	Text exploration map	Culture–unlock task	Multi–path text navigation	Style restoration task	Easter-egg deep guidance
Socializer	AI critique & collaboration	Role-play dialog simulation	AI co-writing task	Scenario-based dialog	Virtual small-group discussion
Killer	AI-versus quiz	Literature challenge match	Beat-the-expert AI	Strategy analysis challenge	Point-capture tournament

11 complete gameplay datasets (including behavior logs, dialogue, and expression sequences) and conducted the annotation under a double-blind protocol. The annotation task encompassed two dimensions: Bartle Player Type: Experts synthesized the player's behavior to categorize them into a single dominant type (Killer, Achiever, Socializer, Explorer). Cultural Cognitive Risk: Experts identified cultural misunderstandings within the interaction and determined the necessity of system intervention. The final benchmark labels were generated via Majority Voting. In cases of significant divergence, the professor adjudicated the final decision, ensuring the authority and reliability of the ground truth.

4.1.2 Ablation Study

To investigate the impact of different modalities and the "Reflection Loop" mechanism on system performance, we established the following five experimental settings for comparative analysis:

MC-MAS (Full): The proposed full model, incorporating multimodal inputs (Expression, Behavior Logs, Text) and the Coordinator's Reflection Loop mechanism with specialized agents (Behavior, Language, and Culture).

w/o Expression: An ablated variant that removes facial expression data, relying solely on text input and behavior logs.

w/o Action: An ablated variant that removes behavior log data, relying solely on text input and facial expressions.

w/o Dialogue: An ablated variant that removes text input, relying solely on behavior logs and facial expressions.

w/o Workflow (No-Reflection): This baseline retains all multimodal inputs but removes the *Coordinator Agent* and its "Reflection–Reconstruction" mechanism. It directly employs a single Large Language Model (`qwen-plus`) to process the multimodal data in a single pass.

Single-agent Self-Reflection: A single-agent baseline where one Large Language Model processes all multimodal inputs and performs one to two rounds of self-reflection using a generic reflection prompt, without agent decomposition or inter-agent coordination.

MC-MAS (No-specialization): A multi-agent baseline that preserves the Coordinator and reflection workflow but removes expert specialization. All agents share identical roles and prompts, differing only in random initialization, to isolate the effect of agent specialization.

4.1.3 Evaluation Metrics

To quantitatively assess the performance on the two core tasks, we define the following metrics:

Bartle Consistency Accuracy (ACC_{Bartle}). Given that the MC-MAS outputs a probability distribution vector P_{agent} over player types,

whereas the human annotation constitutes a single dominant category L_{human}, we employ Top-1 Accuracy for evaluation. A prediction is deemed correct if the player type with the highest probability assigned by the agent matches the human label. The formula is defined as:

$$ACC_{\text{Bartle}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\text{argmax}(P_{\text{agent}}^{(i)}) = L_{\text{human}}^{(i)}) \quad (1)$$

where N denotes the total number of samples, and $\mathbb{I}(\cdot)$ represents the indicator function.

Cultural Attribution Precision (P_{Culture}). This metric evaluates the system's precision in identifying specific "cultural misunderstanding risks." Let S_{model} denote the set of risk points identified by the model, and S_{human} denote the set of valid risk points verified by human experts. Precision is defined as the proportion of model-identified risk points that are validated as "necessary and correct" by the experts:

$$P_{\text{Culture}} = \frac{|S_{\text{model}} \cap S_{\text{human}}|}{|S_{\text{model}}|} \quad (2)$$

4.2 Cognitive Load Evaluation

To evaluate the practical teaching experience of the MC-MAS system in authentic instructional settings, we conducted a controlled user study with 20 participants. The participants were evenly divided into two groups (10 per group). Group A interacted with the Chapter 2 game using the full MC-MAS system with complete multimodal inputs and coordination mechanisms, while Group B completed the same game content using a baseline configuration that removed the MC-MAS framework and relied solely on an end-to-end large language model.

After completing the predefined in-game interaction tasks, participants immediately undertook two subjective evaluation measures to quantify the system's instructional effectiveness and interaction experience.

NASA Task Load Index (NASA-TLX) (Hart and Staveland, 1988). To examine whether the proposed system effectively reduces cognitive barriers in cross-cultural learning scenarios, participants were asked to complete the NASA-TLX questionnaire, which comprises six dimensions, each rated on a 1–100 scale. We focused on three core dimensions: Mental Demand, Frustration, and Perceived Performance. These dimensions were selected to investigate whether the multi-agent collaborative mechanism can maintain an appropriate level of learning challenge while alleviating frustration induced by cultural differences.For the detailed questionnaire, please refer to the Appendix C.

5 Main Results

On the basis of the experimental setup described in the previous section, this section reports the evaluation results of MC-MAS, with a particular focus on the framework's modeling accuracy and user-centered cognitive load.

5.1 Accuracy Evaluation

Table 3 reports the performance of MC-MAS under different ablation settings, evaluated using Bartle player type prediction accuracy (Acc_Bartle) and cultural profile prediction performance (P_Culture). Overall, the full MC-MAS model achieves the best results on both metrics, validating the effectiveness of the proposed multi-modal, multi-agent architecture with a coordinated reflection mechanism.

Impact of Multimodal Inputs. The ablation results show that removing any modality leads to a substantial performance degradation, highlighting the complementary nature of multimodal behavioral signals. Among the three modalities, action logs play a dominant role in Bartle type prediction: removing action data causes Acc_Bartle to drop sharply from 0.912 to 0.455. This indicates that player type inference relies heavily on observable in-game decision patterns rather than solely on expressive or linguistic cues.

In contrast, cultural profile prediction is more sensitive to dialogue data. When the dialogue modality is removed, P_Culture decreases significantly from 0.885 to 0.518, suggesting that cultural traits are more strongly reflected in language use and discourse patterns.

Removing facial expression features results in a moderate but consistent decline across both metrics, indicating that affective cues provide auxiliary information that enhances model performance without being the primary determinant. These findings suggest that MC-MAS benefits from modality specialization, where different input channels contribute unevenly but synergistically to distinct modeling objectives.

Table 3: Performance comparison across different model settings.

Model Setting	Missing Capability	ACC_{Bartle}	Δ	P_{Culture}	Δ
MC-MAS (Full)	–	**0.912**	–	**0.885**	–
Ablation: Modalities					
w/o Expression	Affective cues	0.818	↓0.094	0.754	↓0.131
w/o Action	Behavioral signals	0.455	↓0.457	0.812	↓0.073
w/o Dialogue	Linguistic interaction	0.727	↓0.185	0.518	↓0.367
Ablation: Architecture					
w/o Workflow (No-Reflection)	Iterative reflection	0.758	↓0.154	0.647	↓0.238
Single-agent Self-Reflection	Role diversity	0.805	↓0.107	0.723	↓0.162
MC-MAS (No-specialization)	Expert priors	0.836	↓0.076	0.779	↓0.106

Impact of Architectural Design. Architectural ablations further demonstrate the necessity of multi-agent coordination and structured reflection. Disabling the reflection workflow (*w/o Workflow*) leads to notable drops in both Acc_Bartle and P_Culture, indicating that simple feed-forward aggregation of agent outputs is insufficient for robust inference.

While introducing single-agent self-reflection partially recovers performance, it remains inferior to the full MC-MAS setting, suggesting that self-reflection alone cannot replace inter-agent deliberation. Moreover, the no-specialization variant, in which agents share identical roles, performs better than reflection-free configurations but still falls short of the full model. This highlights that performance gains stem from functional specialization rather than merely increasing the number of agents. For detailed performance comparison, see Appendix Figure 4.

5.2 Cognitive Load Evaluation

As shown in Figure 3, Full MC-MAS consistently outperforms the end-to-end LLM baseline across key NASA-TLX dimensions(measured on a scale of 0 to 100). Participants reported substantially higher perceived performance under Full MC-MAS (71.8 vs. 49.6), indicating improved task effectiveness and confidence. Meanwhile, mental demand was significantly reduced (55.6 vs. 68.9), suggesting that explicit multi-agent coordination alleviates cognitive effort during cross-cultural interpretation. Frustration was also notably lower (36.9 vs. 58.9), reflecting smoother interaction and reduced emotional strain. Overall, these results demonstrate that MC-MAS improves user experience by jointly enhancing perceived performance while reducing cognitive and affective load.

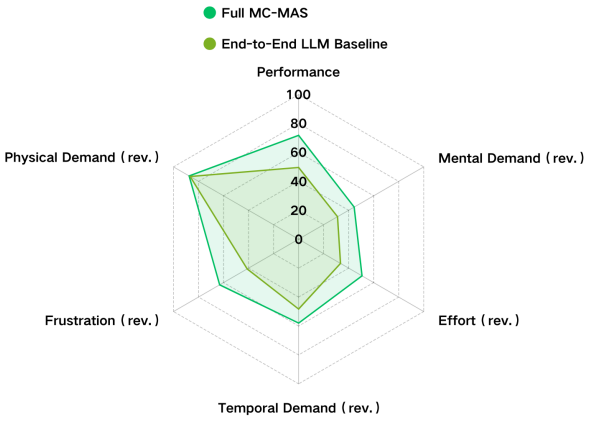


Figure 3: NASA-TLX comparison between Full MC-MAS and End-to-End LLM baseline. Larger areas indicate better performance after direction-aligned normalization.

6 Conclusion

This work examines adaptive narrative generation with LLMs in cross-cultural contexts, showing that multi-agent collaboration effectively enhances semantic and cultural alignment. By coordinating agents modeling linguistic, behavioral, and cultural signals, the framework achieves more accurate user profiling and cultural intention recognition than single-LLM baselines. A reflection-based coordination mechanism further resolves multimodal semantic conflicts and reduces hallucinations, especially in narratives with implicit cultural references. Experiments on a human-labeled Weimar-era literary benchmark validate the effectiveness of the approach. Future work will extend the framework to broader cultural domains and longitudinal user modeling.

Limitations

While the proposed multimodal and multi-agent reflective framework significantly enhances large language models' ability to infer players' cultural backgrounds and gameplay profiles, several limitations remain. One key constraint arises from the deliberate abstraction of input data at the task level. To mitigate hallucination risks associated with long-context processing in large language models, the current system prioritizes semantically summarized interaction states over fine-grained behavioral traces. As a result, detailed low-level signals—such as continuous player movement trajectories, cursor dynamics, or micro-level interaction patterns—are not incorporated into the analytical pipeline. Although this design choice improves model stability and interpretability, it inevitably limits the granularity of player modeling. Incorporating richer behavioral streams in future iterations may enable more comprehensive representations of player intent, engagement, and uncertainty, thereby further strengthening adaptive interventions.

Another limitation concerns the scope of empirical validation. The evaluation system was designed for Chinese learners engaging with German literary content, and the majority of participants shared a relatively homogeneous cultural and linguistic background. While the proposed multi-agent architecture is not intrinsically tied to a specific discipline, language, or cultural setting, its effectiveness in other cross-cultural learning contexts remains empirically unverified. Consequently, the current findings should be interpreted as evidence of methodological feasibility rather than universal generalizability. Future studies involving diverse learner populations and cultural environments are necessary to assess the robustness and transferability of the framework across languages and educational domains.

Ethical Considerations

The use of multimodal data for adaptive player modeling raises important ethical and privacy considerations. In the present system, facial expression information is stored exclusively in a labeled form (e.g., confused, engaged), and no raw facial images are retained. Nevertheless, the collection and analysis of multimodal signals—including language, speech, visual cues, and behavioral patterns—inevitably involve potentially sensitive personal information. Without rigorous data anonymization, access control, and governance mechanisms, such systems may pose risks to user privacy, particularly in educational contexts involving minors or vulnerable populations.

A further ethical concern relates to cultural bias embedded in large language models trained on large-scale corpora. These models may inherit and amplify cultural stereotypes or dominant cultural narratives, which can undermine the goal of fostering genuine cross-cultural understanding. In adaptive educational systems, such biases risk oversimplifying cultural representations or promoting cultural homogenization under the guise of personalization. Balancing cultural authenticity with individualized adaptation therefore remains a critical design challenge.

Looking forward, advances in bias detection and mitigation techniques, combined with participatory design processes involving cross-cultural education experts and target learner communities, are essential for developing more inclusive and culturally responsive systems. Similarly, integrating privacy-preserving data governance frameworks aligned with educational ethics can help reduce risks related to data security and cultural inequity, while maintaining the benefits of multimodal adaptation.

References

Richard A Bartle. 2004. *Designing virtual worlds*. New Riders.

WAL Blyth, BS Bloom, and DR Krathwohl. 1966. Taxonomy of educational objectives. handbook i: Cognitive domaintaxonomy of educational objectives. handbook 2: Affective domain. *British journal of educational studies*, 14(3).

Melania Cabezas-García and Arianne Reimerink. 2022. Cultural context and multimodal knowledge representation: Seeing the forest for the trees. *Frontiers in Psychology*, 13:824932.

Elizabeth B Cloude, Daryn A Dever, Debbie L Hahs-Vaughn, Andrew J Emerson, Roger Azevedo, and James Lester. 2022. Affective dynamics and cognition during game-based learning. *IEEE Transactions on Affective Computing*, 13(4):1705–1717.

Griffin Dietz, Nadin Tamer, Carina Ly, Jimmy K Le, and James A Landay. 2023. Visual storycoder: A multimodal programming environment for children's creation of stories. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Liyan Fan, Jinbo He, Yang Zheng, Yufeng Nie, Taolin Chen, and Hongmei Zhang. 2023. Facial micro-

expression recognition impairment and its relationship with social anxiety in internet gaming disorder. *Current Psychology*, 42(24):21021–21030.

Stanley G Hart and Lowell E Staveland. 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in Psychology*, volume 52, pages 139–183. North-Holland.

Kanwar Sunreep Jossan, Andrea Gauthier, and Jodie Jenkinson. 2021. Cultural implications in the acceptability of game-based learning. *Computers & Education*, 174:104305.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and 1 others. 2024b. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Yishu Li and Yongjian Li. 2025. The paradox of cultural similarity: Teacher adaptation through the lens of linguistic, ethical, cultural, and historical friction. *International Journal of Intercultural Relations*, 109:102295.

Michela Mortara, Chiara Eva Catalano, Francesco Bellotti, Giusy Fiucci, Minica Houry-Panchetti, and Panagiotis Petridis. 2014. Learning cultural heritage by serious games. *Journal of Cultural Heritage*, 15(3):318–325.

Minako O'Hagan. 2009. Towards a cross-cultural game design: an explorative study in understanding the player experience of a localised japanese video game. *The Journal of Specialised Translation*, (11):211–233.

Christopher P Ormell. 1974. Bloom's taxonomy and the objectives of education. *Educational Research*, 17(1):3–18.

Valentina N Pescuma, Dina Serova, Julia Lukassek, Antje Sauermann, Roland Schäfer, Aria Adli, Felix Bildhauer, Markus Egg, Kristina Hülk, Aine Ito, and 1 others. 2023. Situating language register across the ages, languages, modalities, and cultural aspects: Evidence from complementary methods. *Frontiers in psychology*, 13:964658.

Lingzhi Shen, Xiaohao Cai, Yunfei Long, Imran Razzak, Guanming Chen, and Shoaib Jameel. 2025. Calm: Culturally self-aware language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Simone Titus and Dick Ng'ambi. 2023. Digital gaming for cross-cultural learning: Development of a social constructivist game-based learning model at a south african university. *International Journal of Game-Based Learning (IJGBL)*, 13(1):1–20.

Angel Olider Rojas Vistorte, Angel Deroncele-Acosta, Juan Luis Martín Ayala, Angel Barrasa, Caridad López-Granero, and Mariacarla Martí-González. 2024. Integrating artificial intelligence to assess emotions in learning environments: a systematic literature review. *Frontiers in psychology*, 15:1387089.

Abeer A Wafa, Mai M Eldefrawi, and Marwa S Farhan. 2025. Advancing multimodal emotion recognition in big data through prompt engineering and deep adaptive learning. *Journal of Big Data*, 12(1):210.

Hao Wei, Jianing Qiu, Haibao Yu, and Wu Yuan. 2024. Medco: Medical education copilots based on a multi-agent framework. In *European Conference on Computer Vision*, pages 119–135. Springer.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wan, Weihua Luo, and Kaifu Zhang. 2025. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Transactions of the Association for Computational Linguistics*, 13:901–922.

Fang Xie. 2025. From words to multimodalities: Compliment perceptions across lingua cultures. *Journal of Pragmatics*, 239:94–116.

Nan Xie, Zhengxu Li, Haipeng Lu, Wei Pang, Jiayin Song, and Beier Lu. 2025. Msc-trans: A multi-feature-fusion network with encoding structure for student engagement detecting. *IEEE Transactions on Learning Technologies*.

Lijuan Yan, Xiaotao Wu, and Yi Wang. 2025. Student engagement assessment using multimodal deep learning. *PLoS One*, 20(6):e0325377.

Chengcheng Yang, Zhiyao Liang, Tonglai Liu, Zeng Hu, and Dashun Yan. 2025. Mgmr-net: Mamba-guided multimodal reconstruction and fusion network for sentiment analysis with incomplete modalities. *Electronics*, 14(15):3088.

Pardis Sadat Zahraei and Ehsaneddin Asgari. 2025. I am aligned, but with whom? mena values benchmark for evaluating cultural alignment and multilingual bias in llms. *arXiv preprint arXiv:2510.13154*.

He Zhang, Lu Yin, and Hanling Zhang. 2024. Using subjective emotion, facial expression, and gaze direction to evaluate user affective experience and predict preference when playing single-player games. *Ergonomics*, 67(12):1863–1883.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, and 1 others. 2025. Simulating classroom education with llm-empowered agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10364–10379.

A Experimental Setup and Hyperparameters

We provide the detailed implementation settings for our multi-agent experiments in Table 4 to ensure reproducibility.

B prompt

This section illustrate the prompt design for the Behavior Analysis Agent, Language Analysis Agent, Cultural Context Analysis Agent, and Coordinator Agent.

B.1 Behavior Analysis Agent (A_{beh})

Role. The agent acts as a senior game data analyst and behavioral psychology expert, specializing in the Bartle taxonomy of player types.

Input. The agent receives raw player behavior logs $\mathcal{B} = \{r_{\text{c}}, t_{\text{m}}, e_{\text{ratio}}, i_{\text{freq}}\}$ along with the conflict opinion provided by the Committee Leader (S_{Conflict}).

Task. Based on the input data, the agent infers the player's probability distribution over the four Bartle dimensions: $\mathbf{P}_{\text{bartle}} = [\text{Killer}, \text{Achiever}, \text{Socializer}, \text{Explorer}]$.

Theoretical Framework. The inference is grounded in the Bartle taxonomy, where Achievers prioritize in-game accomplishments, Explorers focus on environmental interaction, Socializers emphasize player-to-player interaction, and Killers seek competition and dominance.

Reasoning Anchors.

- **Achiever:** High repetition with increasing success rates, direct trajectories toward task objectives, and resource hoarding behavior.
- **Explorer:** Redundant movement trajectories, frequent low-yield interactions, and unconventional operations.
- **Socializer:** Extended dwell time in public areas, active chat or emoji usage, and a preference for team-based play.
- **Killer:** High PvP participation, preference for burst or control-oriented equipment, and disruptive behavior in non-competitive zones.

Output Format. The agent outputs (1) a normalized probability distribution vector $\mathbf{P}_{\text{bartle}}$ (sum = 100%), and (2) a step-by-step reasoning explanation grounded in observable behavioral evidence such as timestamps, movement trajectories, and interaction heatmaps.

B.2 Language Analysis Agent (A_{lang})

Role. The agent functions as an expert in computational linguistics and affective computing.

Input. The agent simultaneously analyzes the player's dialogue text ($\mathcal{T}$), facial expression data ($\mathcal{E}$), and the conflict opinion provided by the Committee Leader (S_{Conflict}).

Task. The agent performs cross-modal analysis to infer the player's implicit cognitive and emotional states by aligning linguistic features with facial expression cues.

Analysis Focus.

- **Metaphorical Depth:** Examines the use of metaphors or domain-specific game terminology, reflecting the player's level of immersion.
- **Syntactic Complexity:** Assesses whether sentence structures are complex or fragmented, serving as indicators of cognitive state and cognitive load.
- **Emotional Consistency:** Evaluates the alignment between verbal content and facial expressions, including potential discrepancies such as positive language accompanied by negative facial cues (e.g., forced smiling).

Output Format. The agent generates the following outputs, with approximately two sentences for each item:

1. **Player's Linguistic Profile:** A description of the player's contextual depth (high or low) and salient syntactic characteristics.
2. **Player's Emotional State:** A synthesized interpretation of the player's emotional condition derived from cross-modal alignment.

B.3 Cultural Context Analysis Agent (A_{cult})

Role. The agent acts as an expert in cross-cultural communication and a historian specializing in the Weimar Republic period.

Input. The agent analyzes the player's static profile I_{profile} (including nationality and native language), behavioral logs $\mathcal{B} = \{r_{\text{c}}, t_{\text{m}}, e_{\text{ratio}}, i_{\text{freq}}\}$, and the conflict opinion provided by the Committee Leader (S_{Conflict}).

Task. The agent evaluates the cultural distance between the player's background (source culture) and the current in-game context (target culture), with a particular focus on potential cultural misunderstanding risks.

Key Checks.

Table 4: Hyperparameter settings and implementation details for the multi-agent experiments.

Parameter	Value	Description
Model Architecture		
Base Model	Qwen-Plus	Accessed via Alibaba Cloud DashScope API.
Generation Configuration		
Temperature	0.7	Set to encourage diversity while maintaining logic.
Top-p	Default	Using the default setting of the API.
Agent Settings		
Max Dialogue Turns	5	Maximum interaction rounds per session.
System Prompt	See App. B	Role-specific instructions for each agent.

- **Cultural Symbol Identification:** Detects culture-specific references embedded in the current scenario (e.g., biblical allusions or political musical performances).
- **Misunderstanding Prediction:** Predicts potential misinterpretations based on the player's native language and observed behavioral patterns.

Output Format. The agent outputs the cultural assessment result $\mathcal{P}_{\text{Culture}}$, consisting of:

1. **Risk Alert Level** ($O_{\text{RiskAlert}}$): High, Medium, or Low.
2. **Bridging Suggestions** ($O_{\text{BridgingSuggestion}}$): Concrete recommendations aimed at reducing cultural distance and mitigating misunderstanding.

B.4 Coordinator Agent (A_{coord})

Role. The agent acts as the chief architect of adaptive learning, coordinating and supervising a panel of expert agents, including the Behavior, Language, and Cultural Context Agents.

Input. The agent receives analytical reports from all subordinate agents, including detected conflicts (S_{Conflict}) and the current iteration count (N_{iter}).

Task. The agent synthesizes cross-agent reports to determine whether a re-evaluation through a reflection loop is required. If necessary, it specifies the conflict context, triggers re-analysis, and tracks iterative updates until consensus is reached or the maximum iteration limit (5) is exceeded. Upon consensus, the agent generates the final adaptation strategy.

Adjudication Protocols.

- **Conflict Detection:** Identifies semantic or functional inconsistencies across agents (e.g., efficient behavioral patterns conflicting with high cultural misunderstanding risk).
- **Reflection Trigger:** Instructs subordinate agents to re-evaluate their analyses based on the identified conflict context.
- **Decision Priority:** Prioritizes pedagogical efficacy over gameplay progression.

Output Format.

- **If conflict exists:**
  - **Re-evaluation Needed:** Yes.
  - **Conflict Description:** A detailed explanation of the detected cross-modal or cross-agent inconsistency.
- **If consensus is reached:**
  - A final JSON configuration specifying:
    - `Task_Allocation` (adjustment strategy),
    - `Narrative_Style` (adjustment strategy),
    - `AI_Assistant_Tone` (tone setting).

C NASA-TLX Questionnaire

This appendix reports the post-experiment questionnaire used to assess participants' perceived task load after gameplay. The questionnaire is based on the NASA Task Load Index (NASA-TLX) and was adapted to the interactive narrative and cross-cultural learning context of this study.

C.1 Instructions

Participants were asked to rate each item on a continuous scale from 0 (very low) to 100 (very high), according to their subjective experience during the game session.

C.2 NASA-TLX Dimensions

Mental Demand How much mental activity was required to complete the game tasks (e.g., understanding narrative content, interpreting cultural metaphors, selecting dialogue options, and making decisions)?

Physical Demand How much physical effort was required to interact with the system (e.g., mouse clicking and keyboard input)?

Temporal Demand How much time pressure did you experience during interactions with NPCs or while completing tasks in the game?

Performance How successful do you think you were in accomplishing the learning objectives and narrative tasks in the game?

Effort How much mental or physical effort was required to achieve your level of performance?

Frustration How frustrated, stressed, or discouraged did you feel during the game (e.g., due to cultural unfamiliarity or unclear system feedback)?

D Instructions for Expert Evaluators

D.1 Background and Objective

This project aims to validate the accuracy of an automated system in analyzing player game behavior and cultural adaptability. As an expert evaluator, you will be provided with game interaction data collected from 11 anonymous players (including dialogue logs, key decision paths, and behavioral timestamps). Your task is to independently classify each player's characteristics and identify cultural misunderstandings based on this raw data. Your judgments will serve as the **Ground Truth** for the experiment, utilized for comparative validation against the output of the automated workflow.

D.2 Evaluation Tasks

Task 1: Bartle Taxonomy Classification
Please carefully review the player's game behavior patterns to determine which category of the Bartle Taxonomy their dominant motivation best fits. When making a determination, please select the most significant type based on the highest frequency of occurrence or behavior at key decision moments. The specific definitions are as follows:

Achievers Focus on obtaining points, levels, equipment, or other concrete measures of success. Their core motivation is derived from receiving immediate feedback and visible improvement within the game.

Explorers Focus on discovering deep system mechanisms, hidden content, or unknown map areas. They enjoy the process of "figuring out how the game works" more than merely clearing the game.

Socializers Focus on interactions with other players or NPCs. The game itself serves merely as a social backdrop; interpersonal communication and emotional connection are their core motivations.

Killers Focus on imposing influence on others (players or the game environment), typically manifested through competition, destruction, exerting control, or gaining a sense of superiority by dominating the environment.

Task 2: Identification of Cultural Misunderstandings
Please combine the interactive context of the player and the game content to identify moments of cognitive dislocation caused by cultural background differences. The criterion for judgment is: mark instances where the player's behavior or textual feedback clearly indicates a failure to understand the German literary context (e.g., specific social norms of the Weimar Republic era, the intent of Brecht's "Verfremdungseffekt"), or attempts to misinterpret the game situation by applying their own cultural background. Please indicate the specific trigger point (Context) and the nature of the potential misunderstanding in your evaluation report.

D.3 Risk Control and Evaluation Principles

To ensure the objectivity and ethical compliance of the evaluation results, the process must strictly adhere to the following guidelines:

First, regarding **Subjectivity Bias**, given that players may exhibit multiple personality traits simultaneously, please ensure classification is based on **Dominant evidence**, avoiding projections based on personal gaming preferences.

Second, regarding **Over-interpretation**, please maintain a cautious attitude when identifying cultural misunderstandings. Only mark an instance as a cultural misunderstanding when supported by clear evidence (e.g., specific expressions of confusion, repeated ineffective attempts), avoiding the misclassification of common operational errors as cultural barriers.

Finally, all evaluations must be conducted under strict **Independence & Confidentiality**. Please complete the annotation independently without referencing the automated system's results or discussing with other experts to ensure the validity of the blind test. All data is for this research only and is strictly prohibited from external disclosure.

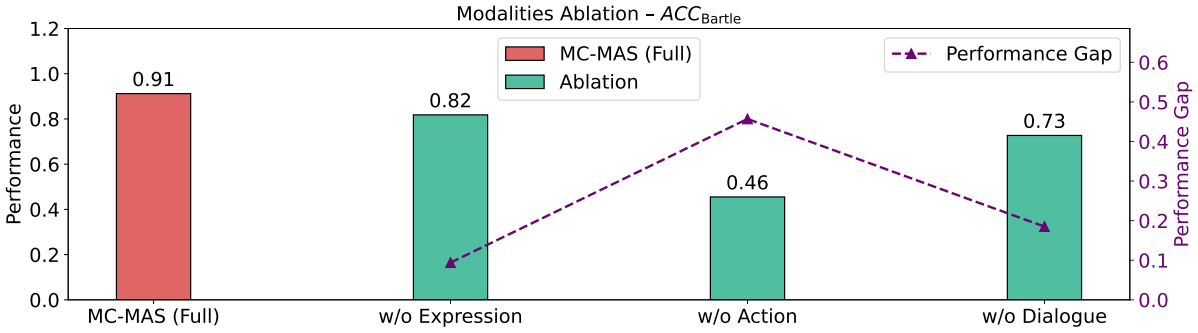
E Recruitment and Compensation Details

In this study, we recruited a total of 20 game participants to engage in the gameplay experiments. Additionally, we invited 11 annotation experts to perform data quality assessment and annotation. All participants were recruited on a voluntary basis.

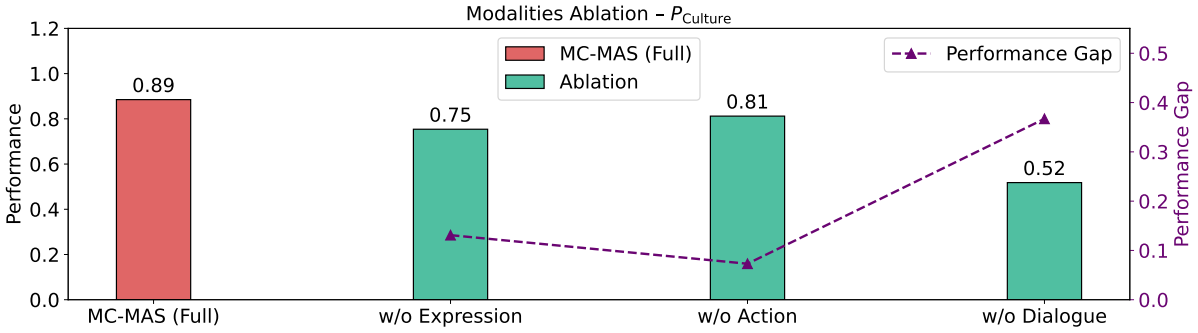
Regarding compensation, each participant (including both game players and experts) received a payment of 20 CNY or a gift of equivalent value upon completion of the task. This compensation plan was designed in accordance with the estimated workload and local ethical standards, and was reviewed and approved by the Institutional Review Board (IRB) of the authors' institution.

F Figure

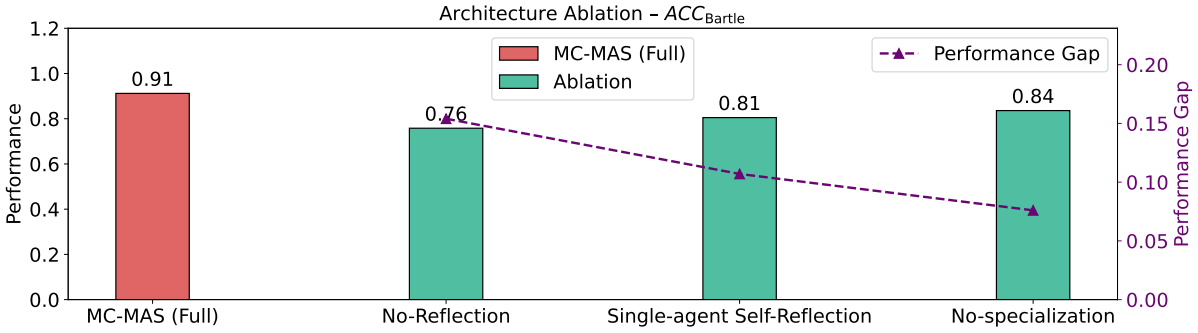
Figure 4 details the ablation study results of the MC-MAS framework on Bartle player type classification ($\mathrm{ACC}_{\mathrm{Bartle}}$) and cultural attribute prediction (P_{Culture}) tasks. The experiment aims to verify the effectiveness of multimodal inputs and core architectural components. The bar charts represent performance scores, while the purple dashed lines illustrate the performance gap between the full model and the ablated variants.



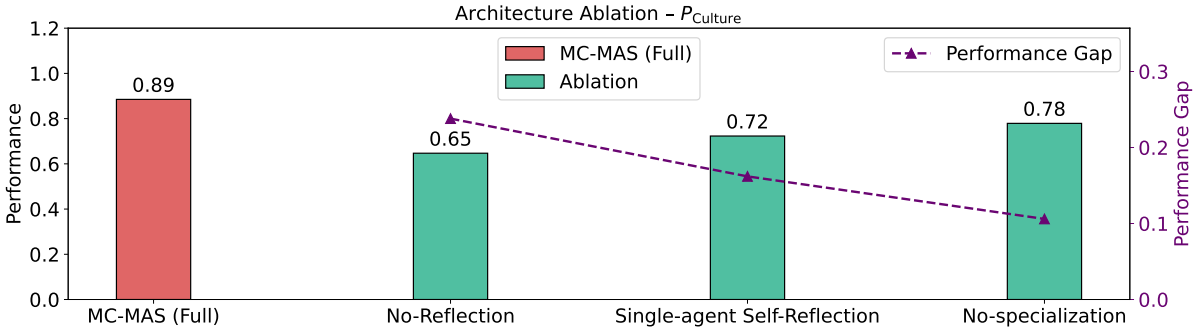
(a) Effect of removing expression, action, or dialogue modalities on Bartle type accuracy



(b) Effect of removing modalities on cultural profile prediction



(c) Effect of architectural modifications on Bartle type accuracy



(d) Effect of architectural modifications on cultural profile prediction

Figure 4: MC-MAS ablation study: Bartle type accuracy and cultural profile prediction with performance gap indicated.