Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.35496

Research on Generative AI Adaptive Interface Regulation Mechanisms for Multimodal Learning Interaction

Yijie Lin

University of the Arts London, London, UK
linyijie02383@outlook.com

Abstract. Multimodal learning environments are now employing more generative AI for explanations, questions, and feedback. Yet, interface designs are still static while rapid changes are observed in user state and cross-modal needs. This paper proposes an adaptive interface regulation framework to coordinate multimodal presentation by using synchronized text, speech, eye movement, and interaction logs. A controlled experiment was conducted on 96 subjects performing reading comprehension, visual interpretation, and interactive questioning tasks using both static and adaptive interface modes. Logs were annotated according to one unified time scheme, and 84 multimodal features were extracted to classify users' states and regulate interface states. Interaction tension and state-to-interface transition were mathematically represented by two functions. Adaptive interface mode decreased mean time taken for each task from 146.3 ± 18.7 s to 127.7 ± 16.4 s, improved average consistency of responses from 0.742 ± 0.028 to 0.851 ± 0.021 ($p < 0.001$), and reduced average time for switching to 1.47 ± 0.31 s from 2.31 ± 0.44 s. Best performance was shown for medium- and high-difficulty scenarios.

Keywords: Generative AI, Adaptive interface, Multimodal interaction, User modeling, Interface regulation

1. Introduction

Multimodal learning environments today involve text, speech, images, and AI-driven dynamic feedback within one interactive process [1]. Here, the learners perform not just reading and listening tasks; they move from one modality to another, comprehend the dynamic changes in output, rephrase prompts, and manage different sources of information in time. AI has expanded the functionality of these multimodal systems through dialogue, explanations, summarization, and guidance through various tasks [2]. However, the interface layer in charge of producing these outputs often does not change, resulting in a mismatch between the dynamic generation of the output and static presentation format [3].

Static interfaces are efficient in situations where task accomplishment and output formation are straightforward. The efficiency becomes low when there is a shift of activities from text reading to viewing, from passive to active information gathering, and from short responses to long-generated answers by the AI. At this point, the user may be faced with challenges such as the difficulty of finding the necessary information, changing screens, and shifting focus across different components

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.35496

of the interface. Problems go beyond visual discomfort; they include low efficiency of interaction, poor mental stability, and low usability of AI services [4].

The key problem is related to the variation in the state of the users over time. While some users continue to perform efficiently, other users suffer from slowdowns, fragmentation, or recovery after mental stress. In such scenarios, if the system is unable to cope with the variation in user states, the output generated by the system might actually increase the level of cognitive friction. Therefore, this research focuses on the regulation of interface adaptively rather than optimizing interface statically. The suggested research methodology involves using interaction data from multiple modes for classifying user state and regulating interface parameters dynamically [5].

2. Literature review

2.1. Multimodal interaction modeling techniques

Multimodal interaction modeling is concerned with the concurrent encoding and decoding of input sources including text, speech, eye-gaze, and interactions with the interface. With respect to learning systems, this entails encoding the dynamics of attention allocation and cognitive processes as well as coordination between behavior and other inputs. The first main problem in multimodal systems is synchronization. That is, because different modes have their own temporal resolution, it is necessary to ensure that events occurring at one mode occur at the same point in time for another mode [6]. Failure to achieve this results in an inability to judge whether the occurrence of change in multiple channels is due to the same interaction event or not. The second issue is feature fusion. This involves early fusion, where inputs are combined from the outset, and late fusion, which retains information about the separate modes before combining them into a decision-making process [7]. Early fusion may be better suited for capturing cross-modal relationships but might suffer more from data sparsity and noise than late fusion.

2.2. Generative AI integration in interactive systems

The use of generative AI brings an added layer of variability to the interaction process in education. Unlike conventional fixed-response systems, the output of generative AI will vary in terms of its length, structure, timing, and even the type of explanations offered, in reaction to the learner's prompts as well as the context [8]. The increased flexibility associated with the system is beneficial. However, the issue becomes problematic with regard to the regulation of such a system. An answer that is relevant in one particular case might be irrelevant in another situation. As a result, the ability of AI to provide assistance effectively will depend both on generation and management of the output interface. The issue will become even more complex in multimodal systems since the generated output will have to be managed together with the source text, visual information, speech input, and navigation tasks. Failure to do this properly will result in increased cross-modal interference [9].

2.3. Adaptive interface design methods

Adaptive interfaces involve efforts to adapt the interface layout, timings, or focus areas based on user activity or demand. Classical approaches use rules, including prolonged periods of inactivity or multiple mistakes made. Modern techniques make use of machine learning in order to predict the current state of a user and adapt the interface. This is especially important in educational contexts, where the goal is to minimize disruption and maximize efficiency [10]. However, despite the

DOI: 10.54254/2755-2721/2026.35496

advances in technology, many adaptive systems still make use of simplistic cues such as number of clicks and/or completion time. These approaches may not be effective for multimodal learning because there are more than one types of cues involved in cases when a user experiences overload, such as unstable gaze, pauses while speaking, and attempts to reformulate their responses. Moreover, it is difficult to regulate the adaptation process itself.

3. Methodology

3.1. Multimodal data acquisition and logging configuration

The framework that was proposed entailed integrating user interaction data drawn from four synchronized streams: text logs, audio signals, eye gaze trajectories, and interface event streams. User interaction logs included prompt length, number of revisions, lexicon variance, pause time, and rate of reformulation. The audio stream involved samples taken at 16 kHz, which provided data on speech duration, proportion of silence, energy variations, hesitations, and spectral change. The eye gaze trajectory data were collected through a 60 Hz eye tracker, and they included fixation duration, saccade amplitude, proportion of gaze dwelling in interface sections, and gaze transition rate. The interface events involved AI response delay, panel switch, interface refresh, and feedback trigger times. All streams were synchronized in time using a common clock system and stored in synchronized time stamps. The regulation demand required developing a multimodal interaction tension function, as shown below in equation (1):

$$\mathscr{R}_t = \sum_{m=1}^{M} \omega_m \hat{z}_{m,t} + \lambda \cdot \frac{\Delta G_t + \Delta A_t + \Delta T_t}{1 + exp(-\eta \tau_t)} \tag{1}$$

where $\hat{z}_{m,t}$ denotes normalized instability in modality m at time window t, ω_m is the contribution weight of that modality, ΔG_t, ΔA_t, and ΔT_t represent short-term changes in gaze dispersion, acoustic fragmentation, and textual reformulation pressure, and τ_t denotes interface switching density. This function combines modality-specific instability with recent interaction turbulence so that the system can estimate whether the current interface configuration is becoming mismatched to the learner's ongoing state.

3.2. Data preprocessing and feature extraction pipeline

Preprocessing of the three streams of information took place independently before their multimodal fusion. The textual data were tokenized and transformed using a pre-trained embedding model. Then, interaction-level features, such as semantic drift in adjacent prompts, correction density, mean complexity of the prompt, and lexical compression rate, were extracted. The audio data underwent denoising and the analysis of their short-time Fourier transform in order to extract spectral centroid, bandwidth, energy variance, voiced-unvoiced ratio, and pause fragmentation. The gaze data were represented using fixation maps, region-based dwell vectors, transition matrices, and attention entropy. If the interval in the gaze data was shorter than 300 milliseconds, then it was interpolated. Otherwise, it was classified as an uncertain segment, and hence ignored during state estimation.

Afterwards, the features were split into 6-second intervals with 50 percent overlap between consecutive intervals. This produced an interaction sequence which had dense temporal information, clear behavioral transitions, and manageable computational burden. Each interval contained 84

DOI: 10.54254/2755-2721/2026.35496

features, including 28 text features, 24 audio features, 22 gaze features, and 10 interface event features. All features were normalized using Min-Max normalization and smoothed at the level of participants to avoid any spikes.

3.3. Adaptive interface regulation model implementation

Adaptive regulation includes two consecutive stages, namely, classification of interaction state and selection of interface configuration. The first stage involves the classification of the corresponding interaction into one of the four possible states, namely, stable-efficient, stable-slow, overloaded-fragmented, and transition-recovery. States should provide a characterization not only of interaction quality but also of the user's need for assistance from an interface. Three classifiers were considered for the task, namely, random forest, multilayer perceptron, and lightweight temporal gated networks. The latter turned out to be most promising in terms of stability-latency-complexity trade-off and thus was selected as the classification component. Classification results were not used directly in regulating interfaces, but rather served as input to the regulation mapping module.

The interface regulation rule was formalized as shown in Equation (2):

$$\mathbf{C}_{t+1} = \arg\max_{\mathbf{c}\in\Omega} \left[\alpha P(U_t \mid X_t) + \beta S(\mathbf{c}, L_t) - \gamma D(\mathbf{c}, \mathbf{C}_t) + \delta H_t\right]$$

where $\mathbf{C}_{t+1}$ denotes the next interface configuration, $P(U_t \mid X_t)$ is the predicted probability of user state U_t given multimodal input X_t, $S(\mathbf{c}, L_t)$ measures the suitability of candidate configuration $\mathbf{c}$ under task difficulty L_t, $D(\mathbf{c}, \mathbf{C}_t)$ penalizes abrupt changes from the current interface state, and H_t represents recent regulation success history.

4. Experimental process

4.1. Experimental task setup and interface configuration

For conducting the experiment, 96 subjects were involved to perform the three multimodal learning activities including reading with AI explanation, interpreting chart graphics, and asking questions through text and speech. In each case, three different difficulty levels of low, moderate, and high were included based on density of information, ambiguity, and need for cross modal coordination. For conducting the experiments, two different interfaces were used. While the static interface retained fixed panel design, output style, and feedback time, in the case of adaptive interface the modulation occurred with respect to focus of contents, density of explanations, modality order, and feedback time according to regulatory logic.

4.2. Data collection procedure and system execution flow

Every participant took part in one static trial and one adaptive trial in a counterbalanced manner. The total time for every trial varied between 24 and 28 minutes. During the whole process, everything related to sensory input, output produced by AI agent, and interactions with the interface was consistently logged. Regulation decisions were made every six seconds during the adaptive trial. Modifications to the interface setup could only be made if the condition had been met during two successive time periods.

DOI: 10.54254/2755-2721/2026.35496

4.3. Model training, validation, and evaluation setup

The data was divided based on users into training, validation, and testing sets according to a ratio of 70/15/15, respectively. Model training used a five-fold cross validation as well as a grid search optimization procedure. User state categorization was measured through weighted accuracy, macro F1 score, and inference delay. Regulation of the interface was determined by assessing time taken, consistency in responses, switching delay, frequency of switching panels, and interaction efficiency.

5. Results

5.1. Performance comparison between static and adaptive interfaces

Clearly, the adaptive interface had significant superiority compared to the static interface (see Table 1). The average task completion time became 127.7 ± 16.4 seconds as compared to the average of 146.3 ± 18.7 seconds. The measure of response variability increased from 0.742 ± 0.028 to 0.851 ± 0.021, while the switch delay was reduced from 2.31 ± 0.44 seconds to 1.47 ± 0.31 seconds. All positive changes were most noticeable when completing difficult tasks – 28.8 ± 4.1 seconds were saved on average

Table 1. Performance comparison between static and adaptive interfaces

Condition / Task	Completion Time (s)	Response Consistency	Interface Switching Latency (s)	Panel Switches per Task	Interaction Efficiency Index
Static Interface, Low Difficulty	118.4 ± 14.2	0.781 ± 0.024	1.92 ± 0.33	4.8 ± 1.1	0.736 ± 0.026
Adaptive Interface, Low Difficulty	110.1 ± 12.7*	0.826 ± 0.022**	1.51 ± 0.28*	4.1 ± 0.9	0.801 ± 0.021**
Static Interface, Medium Difficulty	145.7 ± 16.9	0.744 ± 0.027	2.28 ± 0.39	6.2 ± 1.4	0.691 ± 0.024
Static Interface, High Difficulty	174.8 ± 19.6	0.701 ± 0.031	2.72 ± 0.47	7.3 ± 1.6	0.648 ± 0.029
Adaptive Interface, High Difficulty	146.0 ± 17.5***	0.873 ± 0.018***	1.42 ± 0.26***	5.4 ± 1.1**	0.844 ± 0.017***

5.2. Accuracy of user state classification and adaptation decisions

Weighted accuracy of the classifier was 0.884 ± 0.013 and the macro-F1 score was 0.861 ± 0.016. Stable-efficient category had the best precision, while recovery-transition presented the most difficult task due to similarity to the stable-slow category. The regulation precision was 0.842 ± 0.018, and the regulation recall was 0.817 ± 0.022, which meant that the prediction results for states could be translated to action choices with quite high confidence (Figure 1).

DOI: 10.54254/2755-2721/2026.35496

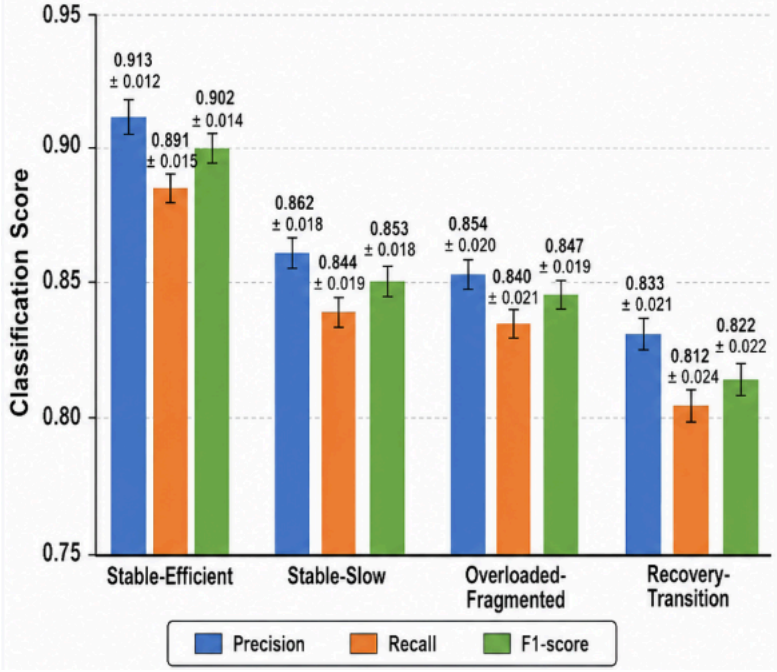

Figure 1. Accuracy of user state classification

5.3. Analysis of different regulation strategies

The strategy level analysis revealed that layout reduction, modal prioritization, and explanation reduction were more beneficial than feedback timing alone. Combination of regulation of all these strategies led to the maximum enhancement, especially in the overloaded-fragmentation state, where interaction efficiency improved by +0.142 ± 0.027*** compared to the control case. From these findings, it can be inferred that disruption in the multimodal interface was usually due to complex mismatches and not a sole mismatch.

Table 2. Accuracy of user state classification and effectiveness of regulation strategies

Metric / Strategy	Stable-Efficient	Stable-Slow	Overloaded-Fragmented	Recovery-Transition	Overall / Gain
Classification Precision	0.913 ± 0.012	0.862 ± 0.018	0.854 ± 0.020	0.833 ± 0.021	0.866 ± 0.015
Classification Recall	0.891 ± 0.015	0.844 ± 0.019	0.840 ± 0.021	0.812 ± 0.024	0.847 ± 0.017
F1-score	0.902 ± 0.014	0.853 ± 0.018	0.847 ± 0.019	0.822 ± 0.022	0.861 ± 0.016
Layout-Density Reduction Gain	—	+0.052 ± 0.016*	+0.081 ± 0.020**	+0.047 ± 0.015*	+0.067 ± 0.017**
Feedback-Timing Adjustment Gain	—	+0.036 ± 0.013	+0.062 ± 0.018*	+0.051 ± 0.015*	+0.049 ± 0.015*
Coordinated Multi-Strategy Gain	—	+0.086 ± 0.019**	+0.142 ± 0.027***	+0.081 ± 0.018**	+0.118 ± 0.021***

DOI: 10.54254/2755-2721/2026.35496

6. Conclusion

An adaptive interface regulation framework for multimodal learning interfaces based on generative AI is introduced in this paper. The use of multimodal synchronization, state detection of the user and configuration of interfaces enables this framework for adaptive interface regulation to facilitate efficient task execution, minimize switch delay and achieve consistent response accuracy when compared with static interfaces. This is especially noticeable in case of task execution in conditions where there is considerable task difficulty. Experiments conducted showed that modeling the user state was critical for interface regulation. It was also found that joint interface regulation outperformed individual interface regulation.

References

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

[2] Xu, W., Wu, Y., & Ouyang, F. (2023). Multimodal learning analytics of collaborative patterns during pair programming in higher education. *International Journal of Educational Technology in Higher Education, 20*, Article 8.

[3] Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: Systematic literature review. *International Journal of Educational Technology in Higher Education, 20*, Article 56.

[4] Ouhaichi, H., Bahtijar, V., & Spikol, D. (2024). Exploring design considerations for multimodal learning analytics systems: An interview study. *Frontiers in Education, 9*, 1356537.

[5] Yan, L., Echeverria, V., Jin, Y., Fernandez-Nieto, G., Zhao, L., Li, X., Alfredo, R., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). Evidence-based multimodal learning analytics for feedback and reflection in collaborative learning. *British Journal of Educational Technology, 55*(5), 1900–1925.

[6] Mohammadi, M., Tajik, E., Martinez-Maldonado, R., Sadiq, S., Tomaszewski, W., & Khosravi, H. (2025). Artificial intelligence in multimodal learning analytics: A systematic literature review. *Computers and Education: Artificial Intelligence, 8*, 100426.

[7] Hariyanto, Kristianingsih, F. X. D., & Maharani, R. (2025). Artificial intelligence in adaptive education: A systematic review of techniques for personalized learning. *Discover Education, 4*, 458.

[8] Feng, S. (2025). Group interaction patterns in generative AI-supported collaborative problem solving: Network analysis of the interactions among students and a GAI chatbot. *British Journal of Educational Technology, 56*(5), 2125–2145.

[9] Chen, D.-L., Aaltonen, K., Lampela, H., & Kujala, J. (2025). The design and implementation of an educational chatbot with personalized adaptive learning features for project management training. *Technology, Knowledge and Learning, 30*, 1047–1072.

[10] Caskurlu, S., Ocak, C., & Dai, C.-P. (2025). The scope of multimodal learning analytics in K–8: A systematic review. *Journal of Learning Analytics, 12*(2), 224–236.