XR Immersive Film and Television Entertainment Interactive Narrative System Based on Multimodal Perception

Junxia Shang[1,2]
[1]Department of Arts, Harbin Conservatory of Music, Harbin 150028, Heilongjiang, China
[2]School of Film and Animation, Chengdu University, Chengdu 610000, Sichuan, China
Corresponding author: Junxia Shang (e-mail: sh33662026@163.com)

ABSTRACT Extended Reality (XR) immersive entertainment systems increasingly rely on real-time multimodal sensing and interactive information transmission to achieve adaptive user experiences. However, existing XR narrative frameworks are primarily constrained by single-modal interaction and limited contextual awareness, resulting in insufficient personalization and delayed narrative adaptation. This study proposes a multimodal perception-driven interactive narrative system that integrates visual streams, speech signals, and physiological sensing through a unified acquisition and temporal synchronization framework. A Transformer-based cross-modal fusion network is employed to jointly encode heterogeneous sensory information and infer user emotions and behavioral intentions, while a reinforcement learning policy dynamically optimizes narrative evolution and cooperates with a large language model for context-aware content generation. An immersive XR feedback loop further coordinates visual rendering, spatial audio propagation, and haptic interaction to establish continuous perception–decision–feedback cycles. Experimental results demonstrate that multimodal emotion recognition maintains an accuracy above 0.80, intention recognition reaches 0.91 during later interaction stages, and users' subjective immersion scores increase from 3.5 to 6.2 with stable low-latency performance. By coupling multimodal sensing, spatial signal propagation, and adaptive interaction within an integrated XR architecture, the proposed framework provides an engineering-oriented paradigm for electromagnetic sensing-enabled immersive environments, intelligent wireless perception, and next-generation human–machine communication systems.

INDEX TERMS XR immersive interaction, multimodal perception, spatial signal propagation, electromagnetic sensing, human–machine communication

I. INTRODUCTION

THE development of extended reality technology has propelled film and television entertainment from linear viewing to immersive experiences, with users' participation in virtual environments gradually shifting from passive reception to active interaction. Existing XR film and television systems largely rely on single interaction methods such as eye tracking and controller operation, limiting their understanding of user behavior to explicit actions while neglecting implicit information such as voice, emotions, and physiological states, resulting in insufficient characterization of user states [1,2]. Narrative logic relies on pre-defined branching structures, lacking in-depth analysis of real-time user feedback, leading to delayed and inconsistent plot adjustments and a significant discrepancy between the immersive experience and individual differences [3,4]. With the development of multimodal perception and intelligent computing technologies, how to integrate multi-source data and drive narrative evolution in XR environments has become an important research direction for enhancing immersive film and television interactive experiences.

Around the XR interactive narrative problem, related research has gradually introduced user behavior analysis and intelligent content generation models. In terms of multimodal perception, Duncan et al. [5] systematically reviewed the multimodal perception methods of human-computer interaction in social environments, providing a theoretical framework for this study to integrate visual, speech and other signals to model user state. Furthermore, in order to realize the transformation from perception to decision, Zhang et al. [6] explored the multimodal interactive perception and decision-making methods based on fuzzy theory in dynamic environments. Their work inspired the design of the decision model of this system when dealing with the uncertainty of user input. In terms of narrative and content generation, Nurjamin et al. [7] verified the effectiveness of multimodal content in improving narrative understanding and immersion by developing and evaluating augmented reality digital narrative videos. This provides a direct reference for this study to construct an XR narrative

environment. Finally, in terms of the specific algorithm of multimodal information fusion, Shao and Wu [8] evaluated and reviewed the multimodal information fusion model in the creation of new media art and film and television culture. Their work laid the methodological foundation for this study to design a cross-modal feature fusion network to drive narrative generation. Some scholars have proposed interaction frameworks based on gaze and gestures to enhance user engagement by modeling user interaction data. Other studies have focused on emotion recognition tasks, introducing facial expression analysis and voice emotion computing into virtual scenes to enhance the system's ability to perceive user states [9,10]. Still other studies have focused on narrative structure design, introducing rule-driven and branching plot control mechanisms to maintain the consistency of story logic [11,12]. These studies have improved the interactive experience to some extent, but they are mostly concentrated on single-modality or weakly integrated strategies, with limited dimensions of user state expression, narrative responses still relying on static structures, and lacking the ability to uniformly model and dynamically adjust complex user intentions [13,14]. With the development of deep learning and cross-modal modeling techniques, multimodal fusion methods have been gradually applied to user behavior understanding and content generation. Related research utilizes attention mechanisms to jointly model visual and speech features, achieving high recognition accuracy in emotion recognition and intent analysis tasks [15,16]. Reinforcement learning methods have been introduced into the decision-making process of interactive systems to adjust system behavior according to environmental states, thereby enhancing interactive adaptability. Large language models have shown strong semantic organization capabilities in text generation and dialogue construction, making dynamic narrative generation possible [17,18]. These methods have shown strong performance in their respective tasks, but the synergistic relationship between cross-modal information in the temporal and semantic dimensions has not yet formed a unified framework; narrative generation lacks tight coupling with user states; and the overall interaction chain of the system exhibits a fragmentation phenomenon [19,20].

This paper focuses on the multimodal perception and dynamic decision-making problems in XR immersive film and television interactive narratives. It constructs a data acquisition and time alignment mechanism that integrates visual, speech, and physiological signals to form a unified multimodal input representation. A Transformer-based cross-modal fusion model is introduced to jointly encode multi-source features and generate user emotion and intent representation vectors. At the narrative layer, a reinforcement learning-driven decision-making model is introduced to map user states to the plot evolution path, and a large language model is combined to complete narrative content generation and semantic expansion. At the system layer, an immersive feedback mechanism is constructed to form a closed-loop structure between perception analysis and narrative adjustment. This research offers new insights into the depth of multimodal information fusion, dynamic narrative adjustment methods, and the construction of interactive closed loops, providing a new technical path for the intelligent development of XR film and television entertainment systems.

II. METHODS

A. MULTI-SOURCE SENSING DATA ACQUISITION AND SYNCHRONIZATION MECHANISM

This study constructs a multimodal data acquisition system to uniformly acquire multidimensional behavioral information of users in the XR environment. Visual data acquisition adopts the collaborative work of depth camera and RGB camera, uses the facial key point detection algorithm to extract facial expression features, combines the pose estimation model to analyze the movement trajectory of head and upper limb, and outputs continuous time series data at a fixed frame rate [21]. Speech signal acquisition relies on array microphone device to acquire user speech input, and after removing silent segments through endpoint detection algorithm, it is input into speech recognition model, converts speech stream into text sequence and outputs semantic embedding vector synchronously. Physiological signal acquisition relies on wearable device to record heart rate and skin conductance changes in real time, and uses filtering algorithm to remove high frequency noise and motion artifacts, mapping continuous physiological data into stable temporal features. A unified time synchronization mechanism has been developed to solve the problem of inconsistent sampling frequencies between multiple sources of data. A high-precision timestamping module is attached to the data acquisition side, allowing all visual frames, pieces of speech, and pieces of physiological signals to be labelled uniformly. To support the differences in the frame level and segment level of visual and audio data, a sliding-window resampling method has been used so that each modality can be represented using a common time scale. For physiological signals, which are characterized by high frequency sampling and biological latency, a time-delay compensation alignment mechanism is applied in addition to piecewise averaging and interpolation. Specifically, a dynamic lag adjustment (approx. 1.5-3 seconds) is introduced to synchronize the delayed physiological responses (e.g., heart rate) with the instantaneous visual and speech cues, ensuring that the multimodal features reflect the same affective state.

During data encoding, a single feature representation structure has been created whereby all visual expression, pose vector, semantic embedding, and physiological signal features have been standardised. A normalisation algorithm is then applied to the data to remove any dimensional differences. Finally, a vector concatenation method has been used to create a multimodal joint representation. During the encoding process, a masking mechanism is used to identify and mark the missing or abnormal segments of data to

prevent redundancy and interference between modalities and ensure the reliability of the fusion process that follows.

B. CROSS-MODAL FEATURE FUSION AND USER STATE MODELING

After completing the temporal alignment and unified encoding of multi-source data, a cross-modal feature fusion network is constructed to achieve joint modeling of multi-dimensional information. First, visual expression vectors, pose features, semantic embeddings, and physiological signal sequences are respectively input into a linear mapping layer to project features of different dimensions onto a unified dimensional space, obtaining an initial modal representation [22,23]. To eliminate the distribution differences between modalities, a layer normalization operation is introduced to standardize the features of each modality, and positional encoding is added at the input to preserve the temporal relationship in the time series. Subsequently, the features of each modality are concatenated according to time steps to form a multimodal sequence, which is then input into the cross-modal Transformer encoder.

During the encoding process, a multi-head self-attention mechanism is used to assign weights to features of different modalities, enabling the model to adaptively select discriminative information at each time step. The attention calculation process is defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

A linear transformation, applied to the input sequence, produces query vectors (Q), key vectors (K), and value vectors (V). The attention weight matrix reflects the relationship between different modalities; therefore, the cross-modal attention mechanism enables cross-modal information to interact and is enhanced by the shared features of the two or more modalities. To further increase the semantic alignment ability of the two or more modalities, an additional encoder layer using cross-modal attention is added. Therefore, visual features act as Qs, while speech and physiological features are K-V pairs for updating each other interactively to create a stronger association between user behaviours and emotions.

In the feature fusion output stage, the encoded multimodal representation is input into the gated fusion unit to dynamically adjust the contribution of different modalities, suppress redundant information and highlight key features. The fusion result is input into the emotion recognition subnetwork, which uses a multilayer perceptron structure to perform nonlinear mapping on high-dimensional features, outputs discrete emotion category probability distributions, and simultaneously calculates continuous emotion intensity values [24,25]. The emotion prediction process is represented as follows:

$$y = \text{softmax}(Wf + b) \quad (2)$$

The fused feature vector is used as input, and the probability of the emotion category is obtained after being mapped by the weight matrix.

Figure 1 illustrates the changes in the distribution of emotion features in t-distributed stochastic neighbor embedding (t-SNE) before and after cross-modal feature fusion. In the image before fusion, the spatial distribution of each emotion category (Neutral, Happy, Sad, Angry, Fear, Surprise, Disgust) highly overlaps, and the category boundaries are blurred. This indicates that visual, speech, and physiological features have strong feature redundancy and inconsistent expression in the unfused state, making it difficult to form a stable discriminative structure. In the image after fusion, different emotion categories show a clear clustering trend in two-dimensional space, and the interval between each category is significantly increased. This indicates that the cross-modal features, under the action of the Transformer fusion mechanism, enhance the inter-class separability and structural consistency.

In the behavioral intent modeling stage, an intent reasoning network based on sequence modeling is constructed. The fused features from the temporal dimension are input into a bidirectional encoding structure to perform contextual modeling of the user's historical behavior and output the intent label for the current moment. To enhance inference stability, a cross-entropy loss function is introduced during the training phase to constrain the emotion classification results, and sequence consistency constraints are combined to optimize the intent prediction process. Finally, a user state representation containing an emotion state vector and a behavioral intent label is formed. This representation maintains continuity and consistency in the temporal dimension, providing structured input for the subsequent narrative decision-making module.

C. ADAPTIVE INTERACTIVE NARRATIVE GENERATION MECHANISM

The narrative decision-making model using reinforcement learning is based on a user state vector. The user's emotional state vector and label for a behavioral intention are defined together as the environmental state, while the current plot node and its optional branches are encoded using a discrete action space. A reward function is then created based on narrative coherence and consistency with user feedback as optimization objectives.

As depicted in Figure 2, the x-axis represents the training episode of the narrative decision-making strategy, while the y-axis represents the cumulative reward of the narrative generation strategy throughout training. The data shown illustrates the trend of convergence of the narrative generation strategy throughout interactive learning. The curve's cumulative reward at the beginning of training contains large swings in the value of rewards, demonstrating that the policy network's choice of plot branches is very exploratory, but also very unstable in the early stages; therefore, most of the reward values are low and contain a lot of random

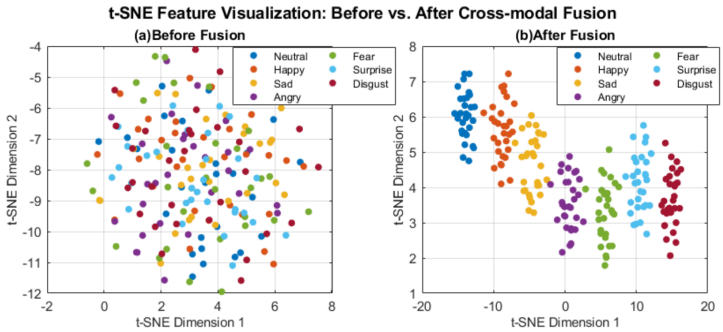

FIGURE 1. Comparison of t-SNE distribution before and after cross-modal feature fusion: (a) t-SNE distribution of high-dimensional features before fusion; (b) t-SNE distribution of high-dimensional features after fusion

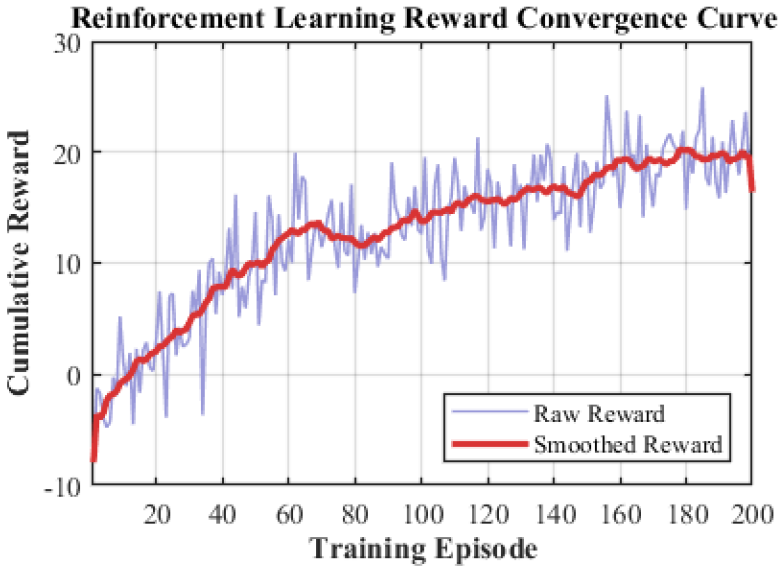

FIGURE 2. Reinforcement learning reward convergence curve

noise fluctuations. The cumulative reward curve gradually increases and stabilizes after around 160 episodes, indicating that as there are more training episodes, the policy network learns to select better narrative selection paths that improve the matching degree between plot branch selection and the user state. Furthermore, in the last stage where the cumulative reward is stable at a larger value, this indicates that the reinforcement learning model has learned to develop a better policy distribution in the decision space of the narrative, and thus can schedule plots much more reasonably for different user states.

The state vector is input into the policy network, and a deep neural network is used to map the high-dimensional state, outputting the selection probability distribution of each plot branch. The current optimal narrative path is selected through sampling or the maximum probability policy [26,27]. The policy update process iterates the parameters according to the policy gradient method, and its objective function is expressed as:

$$J(\theta) = E_{\pi_\theta}\left[\sum_t r_t\right] \tag{3}$$

The policy function is controlled by parameter θ, and the cumulative reward reflects the degree of matching between the plot choice and the user's state. To improve training stability, a baseline function is introduced during the policy update process to normalize the reward, reduce gradient estimation variance, and make the narrative path selection process smoother.

At the plot generation content stage, the current user state and the RL output plot nodes are input conditionally into a large language model, which encodes a narrative context representation through a context encoding process. An autoregressive approach to language generation is used by the large language model through the gradual construction of the generated dialogue and plot description text from the current plot context, as well as by implementing a strict prompt-based state-transition constraint during content generation. The output action selected by the Reinforcement Learning (RL) model serves as a mandatory prefix and semantic boundary for the LLM. This hard constraint prevents the LLM from "hallucinating" details that contradict the RL-selected branch, ensuring the generated narrative remains strictly within the logical path of the decision-making model

[28,29]. The discrete action output from the reinforcement learning model serves as a semantic boundary and prefix constraint for the large language model. By injecting the RL-decided plot node into the prompt as a mandatory logical anchor, the system effectively prevents the LLM from hallucinating narrative details that contradict the chosen branch. A conditional language model defines the generation probability distribution.

$$P(y \mid x) = \prod_{t=1}^{T} P(y_t \mid y_{<t}, x) \tag{4}$$

The input variables include plot node information and user state representation, and the output sequence corresponds to the generated dialogue and narrative text.

A contextual memory unit is developed throughout the creation phase of the narrative in order to encode and archive historical plot points of the previous generating phases and the generated text; this contextual memory unit serves as supplemental input that assists in creating semantic models when generating future content, and also serves to prevent logical breaks and semantic conflicts between future decision-making processes and previous generated content. A circular feedback loop has been established linking narrative decision making with content creation by re-mapping future user actions and emotional transitions to the state space and dynamically correcting the rewards functions so as to continuously adjust the plot choice tendencies in the policy network. Therefore, through the establishment of this user state driven narrative generation chain, it can be ensured that there is continued consistency between the selection of a plot branch and its associated content generated over time, thereby greatly increasing the adaptability or rhythmic coherence of the interactive narrative creation process.

D. IMPLEMENTATION OF XR IMMERSIVE FEEDBACK AND INTERACTION LOOP

After the narrative decision and content generation are completed, the plot nodes and generated text are mapped to the scene control parameters in the XR engine to realize multichannel immersive feedback scheduling. First, the narrative output is structured and parsed, and the plot tags, emotional states and semantic information are converted into scene control instructions, which correspond to light intensity, color temperature change, sound effect type and character action sequence respectively [30]. The visual feedback module dynamically updates the scene light and shadow based on the real-time rendering engine, and uses the physical basic rendering model to adjust the light source parameters so that the ambient brightness and color distribution are linked with the changes in narrative emotion. The audio feedback module selects the corresponding sound effect resources according to the semantic tags, and calculates the sound source position and propagation path through the spatial audio algorithm so that the sound direction is consistent with the user's perspective, thereby enhancing the sense of spatial immersion. The narrative generation aspect includes converting narrative generation results into a collective of character action instructions. The model for controlling character action is based on an animated skeletal model, with the presented facial expression adjustment being performed algorithmically and designed to synchronize the character's facial expression with the character's narrative. For the haptic feedback component, the system obtains wearable instructions indicating how to map the narrative's emotional context into the vibration frequency and intensity signals that generate a tactile stimulant applied to the user to enhance the multisensory interactivity of the experience. In order to maintain consistency among the multiple feedback elements in the time domain, the creation of a synchronization scheduling mechanism is integrated into the system for the purpose of uniformly scheduling and delaying audio, visual, and haptic signals together. The use of a buffer queue provides a mechanism to align the outputs of each channel in order to minimize the time offset between the different multi-channel feedback received by the user.

In terms of constructing the interactive closed loop, the user's behavioral and physiological changes during the feedback phase are re-input into the multimodal perception module, forming a continuous data stream. The system performs state updates at fixed time steps, transmits the perception results to the narrative decision module, and recalculates the plot branch choices and content generation results based on the latest state. The closed-loop process satisfies the state transition relationship:

$$s_{t+1} = f(s_t, a_t) \tag{5}$$

The current state and executed action determine the system state at the next moment. Asynchronous processing has been implemented in the closed-loop system to improve response time of subsystems. The perception, decision-making, and rendering processes are executed along independent threads using shared memory for data exchange and latency reduction. Through continuous iterations and feedback scheduling, the stable time dimension operating chain of the system is established between user input, system decisions, and environmental feedback, thereby ensuring continuity and consistency of XR immersive interaction.

III. RESULTS

A. DATASET SOURCE AND EXPERIMENTAL CONFIGURATION

The current research project incorporates two sources of existing multimodal interactive data, as well as an original source. It uses the publicly available RAVDESS dataset to create a sample of speech emotions, and it contains 1,440 instances of speech with 7 distinct emotions in total, recorded at 48kHz. It uses FER2013 to produce a sample of facial expressions and contains 35,887 images with 7 distinct basic facial expressions, each 48×48 pixels in size and recorded in grayscale. Finally, it uses DEAP to create a

sample for the categorization of physiological states, which consists of measurements of EEG signals from 32 subjects with 32 EEG signals recorded at a frequency of 128Hz and approximately 60 seconds each [31,32].

In the XR interaction scenario section, an extended experimental subset is constructed. The publicly available data and user behavior simulation data are time-aligned and recombined to form multimodal sequence samples. Each sample contains a visual frame sequence, a speech segment, and a synchronized physiological signal segment. The time window is set to a 2-second sliding sampling with a step size of 0.5 seconds. The training set and the test set are divided in an 8:2 ratio, with approximately 28,000 multimodal sequences for training and approximately 7,000 sequences for testing.

For the experimental settings, an NVIDIA RTX 3090 GPU with 24GB of video memory, CUDA version 11.7, and the PyTorch 2.0 deep learning framework are used as the basis for generating the dataset. A fixed-sized input is used as the model of a 512-dimensional feature vector; 6 transformer encoding layers of computation are set up; and 8 multi-head attention heads are set up for processing during the inference and training procedures. The batch size for testing is a maximum of 32. The Adam optimizer is used as the learning algorithm because it provides fast convergence and is easy to adjust. A linear decay approach [33] is also used to automatically reduce the initial learning rate (1e-4) as progressing through the dataset. The multimodal processing consists of visual and audio (speech) signals normalized together, and physiological signal state is denoised using a bandpass filter and is uniformly resampled to a constant time step to facilitate comparison across all modalities.

B. CONSISTENCY ASSESSMENT OF IMMERSIVE EXPERIENCE

Participants are chosen to take part in a shared experience (XR) or an interactive story within a single XR scenario. User activity (time spent, location of gaze path and frequency of interactions) can be documented at multiple plot points throughout the duration of the experiment. Following the experiment, participants complete an immersion scale that assesses the level of spatial presence, engagement and emotional engage ment they feel. The correlation between participant's behavioural data and immersion scale ratings can then be determined. Each user's behavioural fluctuations over time are analysed statistically by segments. The mean and variance of each rating can also be calculated. The immersion performance of the experience during each stage of the narrative is compared and analysed in relation to both the rating distributions and trends evidenced through participant's behavioural data in relation to the participants, for overall experience stability under fluctuating feedback.

The data in Figure 3 helps to evaluate how consistent; each user reacts to and experiences the immersive film and television while interacting with XR. Figure 3(a) is a graphical representation of user dwell time (in seconds) across the horizontal (X-axis) (Plot Node - Intro, Dev1, Conflict, Climax, Dev2, Ending) and vertical (Yaxis) (User Dwell Times, in seconds, at each Plot Node). Figure 3(a) provides two different methods (Boxplot and Scatterplot) for visualizing user data collection results that show how much variance exists across all users in terms of their ability to maintain focus (or lack thereof) based on the length of time that they spent at each of the five plot nodes during the Climax plot node (highest median value - Climax node > 50 seconds) is peak for user attentiveness behaviors. The dwell time at the Intro and Dev1 stages is lower and fluctuates less, reflecting that the interaction load is lower in the early stage of the narrative. Figure 3(b) shows that the horizontal axis represents the length of the gaze trajectory (pixels), and the vertical axis represents the interaction frequency. The scatter plots are grouped according to the level of immersion. The high immersion group is mainly distributed in the high trajectory length and high interaction frequency areas, indicating that the expansion of visual attention and the enhancement of interaction behavior are positively correlated. The linear fitting curve shows that the two are generally positively correlated, and the correlation coefficient is high, indicating that gaze behavior has a strong explanatory power for interaction intensity. This result verifies that there is a consistent mapping relationship between multimodal behavioral data and subjective immersive experience.

The temporal evolution of narrative interaction is illustrated in Figure 4 along the horizontal axis; each time point represents a specific two-minute interval of XR immersive film and TV interactive experience. The left half of the vertical axis shows standardized frequency for user interactions, while the right half shows an immersion score from a Likert scale (1-7). As indicated by the behavioral curve, the frequency of user interactions shows a slow gradual increase until it reaches a peak at time points 4-6, and then slowly starts to decrease, indicating an increase in the rate of user-interaction behaviors at the peak moment (climax) of the plot and subsequently a decrease in rate of user-interaction behaviors once users have entered into the conclusion of the plot. The subjective score curve increases from 3.5 to 6.2 during the same time period, indicating a synchronous relationship between immersion and the intensity of user-interaction behavior and a relatively stable decline following the peak moment (climax) of narrative engagement. This demonstrates that the multimodal perception-based narrative mechanism significantly enhances users' overall experiences during key plot development points and maintains an overall congruency of both behavioral data and subjective data, demonstrating the systematic coupling of user-interaction behavior and subjective experiences resulting from perceptual feedback.

C. NARRATIVE RESPONSE AND INTERACTION TIMELINESS ASSESSMENT

When the system is operating, time data is taken from the user input to the system's output in real time, generating

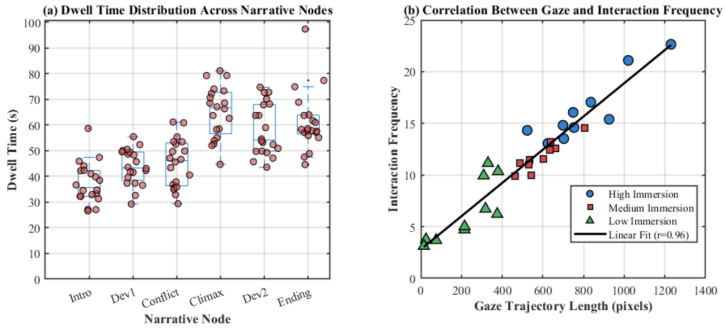

FIGURE 3. Immersive experience consistency assessment chart: (a) Distribution of user dwell time at different story nodes; (b) Correlation analysis chart of eye tracking and interaction frequency

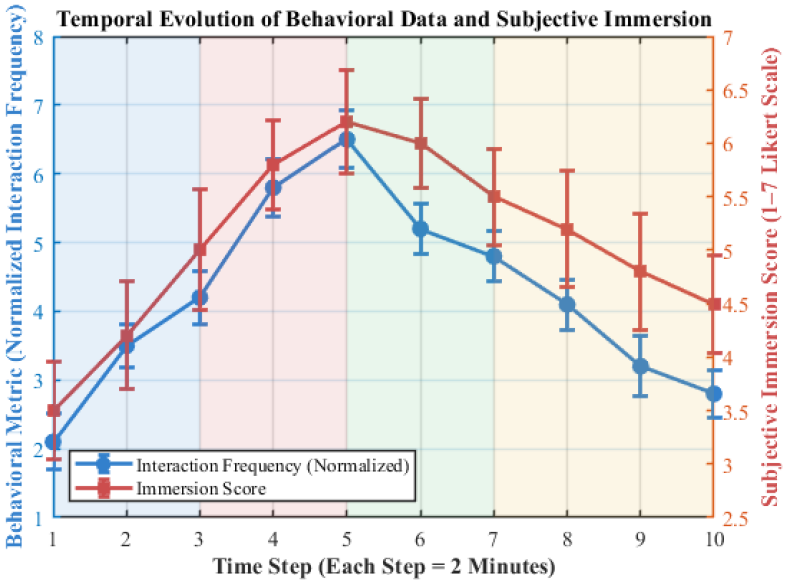

FIGURE 4. Relationship between behavior and subjective immersion over time

time data on the three processes comprising the multimodal perception module's processing time, the narrative decision generation time, and the XR rendering feedback time. For each process, the time data collected is divided into time series according to the timeframe for each type of response latency as measured within each interaction cycle; therefore, the processing times for other modules can also be recorded separately. Following the collection of stable data from multiple experimental conditions, the average response time for the experimental conditions is determined and evaluated for variability; furthermore, the evaluation of timeframe data from various scenarios can then be used to evaluate the response stability and processing efficiency of the system under conditions of frequency of interaction.

Table 1 displays the processing times for each of the multimodal modules calculated at low, moderate, and high complexity levels. As complexity increases, the average processing time of the multimodal perception module (42.3 ms to 56.2 ms), narrative decision module (115.6 ms to 146.9 ms), and XR rendering feedback module (78.5 ms to 96.4 ms) also increases. The total interaction latency at increased levels of complexity increases from 236.4ms to 299.5ms respectively. The maximum latency for the narrative decision-making module has the largest percentage for processing (approximately 49%) of the aggregate processing time in the highest frequency interaction scenario based on complexity level although the maximum total latency experienced all modules collectively is managed within 407ms. To mitigate the perceived sluggishness in the high-complexity scenario (299.5ms), the LLM inference is optimized using 4-bit quantization and KV caching, while the Transformer encoder utilizes a sparse attention mechanism to reduce computational overhead. These optimizations prioritize maintaining the motion-to-photon latency within acceptable bounds for immersive interaction. To achieve this under high-complexity scenarios, the LLM utilizes a 4-bit quantization and KV cache optimization, and the 6-layer Transformer employs model pruning to reduce redundant attention computations. These optimizations ensure the total interaction latency stays below the 300ms threshold, minimizing the perceptual "sluggishness" for users. Standard deviations for each of the modules processing times also expand as complexity increases, demonstrating the increased volatility present in a system at maximum

TABLE 1. Statistics on processing time of each module in the system

Module	Scenario Complexity	Mean Latency (ms)	Std (ms)	Max Latency (ms)
Multimodal Perception	Low	42.3	5.1	58
Multimodal Perception	Medium	48.7	6.4	69
Multimodal Perception	High	56.2	8.3	84
Narrative Decision	Low	115.6	12.2	148
Narrative Decision	Medium	128.4	14.5	167
Narrative Decision	High	146.9	18.1	192
XR Rendering Feedback	Low	78.5	7.8	99
XR Rendering Feedback	Medium	85.2	9.3	110
XR Rendering Feedback	High	96.4	11.6	131
Total Interaction Latency	Low	236.4	18.2	305
Total Interaction Latency	Medium	262.3	21.4	346
Total Interaction Latency	High	299.5	26.9	407

load conditions. However, the aggregate processing time increase corresponding to increasing complexity levels are fairly moderate in comparison to the maximum aggregate processing time increase of approximately 27% from low to high complexity levels, thus supporting the systems stability of response time and efficiency of processing in relation to high frequency interactions.

D. USER STATE RECOGNITION AND NARRATIVE MATCHING EVALUATION

The emotions expressed by users, along with those predicted by the system, are recorded as they occur in real-time. These expressions (facial and vocal) and physiological signals, are compared against manually categorised emotions and behavioural intentions. The emotion label generated by the system is then compared against the user's intention recognitions on a frame-by-frame basis, with the number of correct/incorrect classifications and recognition breaks also being counted based on time segmentation. Following this, further data is analysed to assess the recognition deviation from the actual narrative points. The recognition data is then combined with the narrative plot selection paths to carry out a statistical analysis between the paths proposed by the system with those chosen by the user. This analysis helps determine whether or not user state modelling supports decision-making in a narrative environment.

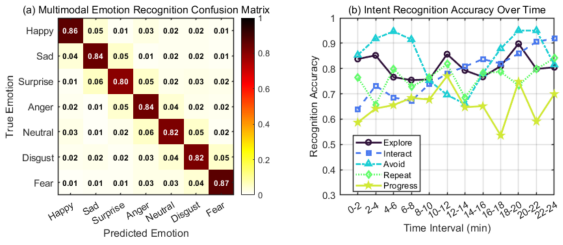

FIGURE 5. Performance evaluation results of user state recognition: (a) Confusion matrix of multimodal emotion recognition; (b) Change in intent recognition accuracy over time

The capability of the system to perceive a user's emotional state is assessed at two levels: identifying the user's emotion and understanding their intended action. In Figure 5(a), the confusion matrix for emotion recognition is presented with the predicted emotions on the horizontal axis and actual emotions represented on the vertical axis. The intensity of the color indicates the probabilities associated with each classification. It can be noted that most values of the confusion matrix are high (above 0.80), which indicates that the model has a significant level of fidelity in identifying the major emotion categories. In Figure 5(b), the accuracy of intent recognition by the system is shown over time. The time axis is represented by the period of interaction in minutes (0-24). Each of the curves represents a different type of anticipated action (e.g., Explore, Interact, Avoid). Overall, the system maintains an average recognition rate of greater than or equal to 0.70. The intent to "Interact" also demonstrates an increase, reaching greater than or equal to 0.91 by the latter half of the period measured and the intent to "Avoid" experiences some variation. Therefore, the ability of the system to adapt to dynamic behavior increases as the narrative continues. In addition, the improvement in model performance at significant time points is indicative of how well the model is able to accurately represent changes to the context of the narrative.

IV. CONCLUSION

The current study examines XR immersive film and TV interactive narrative, developing a narrative generation system through utilizing of multimodal and perception based modelling with visual, auditory and physiological signalling. The system employs Transformer for modal cross-fusion and reinforcement learning to determine user state, thereby enabling plot branching and real-time mapping of user state to plot branches. A large language model has also been employed to generate the narrative content for creating a feedback loop of perception to decisions made, narrative generation and feedback. The results of the experiments show consistent performance of the system across all tests, providing evidence of the effectiveness of the combination of multimodal fusion and dynamic narratives. The major drawback of the system is that it lacks the capability for generalising across scenes and requires a substantial amount computing power; future designs may wish to pursue more efficient and compact model structures and adaptive training of multi-scene.

AUTHOR CONTRIBUTIONS

Junxia Shang designed, collected and analyzed the data, and drafted the manuscript. Junxia Shang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Junxia Shang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This article is the preliminary research result of the 2025 project 'Practical Research on Aesthetic Education in Primary Schools in Longquanyi District, Chengdu from the Perspective of 'Five Educations Together'' (TC- CXJY-2025-B13) of the Urban and Rural Education Development Research Center under the Key Research Base of Humanities and Social Sciences in Higher Education Institutions of Sichuan Province.

Chengdu University 2024-2026 Teaching Reform and Practice Project: Teaching Reform and Practice of the Applied Urban University Public Art Course 'Art Appreciation', Project Number: cdjgb2024201

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] H. Wang, Z. Feng, X. Yang, et al., "MRLab: Virtual-reality fusion smart laboratory based on multimodal fusion," International Journal of Human-Computer Interaction, vol. 40, no. 8, pp. 1975-1988, 2024, doi: 10.1080/10447318.2023.2227823.

[2] C. Chen, K.ZK. Zhang, Z. Chu, et al., "Augmented reality in the metaverse market: the role of multimodal sensory interaction," Internet Research, vol. 34, no. 1, pp. 9-38, 2024, doi: 10.1108/INTR-08-2022-0670.

[3] B. Hussain, J. Guo, F. Sidra, et al., "Enhancing spatial awareness via multimodal fusion of cnn-based visual and depth features," International Journal of Ethical AI Application, vol. 1, no. 3, pp. 13-27, 2025, doi: 10.64229/gdz8tc37.

[4] L. Cao, H. Zhang, C. Peng, et al., "Real-time multimodal interaction in virtual reality-a case study with a large virtu-al interface," Multimedia Tools and Applications, vol. 82, no. 16, pp. 25427-25448, 2023, doi: 10.1007/s11042-023-14381-6.

[5] J.A. Duncan, F. Alambeigi, and W. Pryor M, "A survey of multimodal perception methods for human-robot interaction in social environments," ACM Transactions on Human-Robot Interaction, vol. 13, no. 4, pp. 1-50, 2024, doi: 10.1145/3657030.

[6] J. Zhang, S. Wang, W. He, et al., "Perception and Decision-Making for Multi Modal Interaction Based on Fuzzy Theory in the Dynamic Environment," International Journal of Human-Computer Interaction, vol. 40, no. 24, pp. 8794-8808, 2024, doi: 10.1080/10447318.2023.2291607.

[7] A. Nurjamin, L. R. Nurjamin, Y.N. Fajriah, et al., "Developing and evaluating an augmented reality (AR) digital storytelling video to foster multimodal literacy and narrative comprehension," Journal of Engineering Science and Technology, vol. 20, no. 4, pp. 919-956, 2025.

[8] J. Shao and D. Wu, "Evaluation on algorithms and models for multimodal information fusion and evaluation in new media art and film and television cultural creation," Journal of Computational Methods in Science and Engineering, vol. 24, no. 4-5, pp. 3173-3189, 2024, doi: 10.3233/JCM-247565.

[9] Y. Ji, "Artistic alchemy: Exploring the fusion of art theory and film aesthetics in visual storytelling," Herança, vol. 7, no. 2, pp. 51-68, 2024, doi: 10.52152/heranca.v7i2.785.

[10] S. M. Jamasbi and A. Ghazvineh, "A Multimodal Analysis of Films: Toward a Peircean Framework," Quarterly Review of Film and Video, vol. 42, no. 6, pp. 1543-1565, 2025, doi: 10.1080/10509208.2023.2265783.

[11] Z. Cai and K. Liu, "Construction of interactive narrative in children's drama driven by generative adversarial networks," International Journal of Information and Communication Technology, vol. 27, no. 30, pp. 1-23, 2026, doi: 10.1504/ IJICT.2026.152655.

[12] A. Raheel, D. Khalid, and S. S. Ahmed, "Classifying Emotions in 3-D: Physiological Insights Into Tactile-Audio-Visual Immersion," IEEE Sensors Journal, vol. 26, no. 2, pp. 2535-2542, 2025, doi: 10.1109/JSEN.2025.3637217.

[13] P. Qi, "Movie visual and speech analysis through multimodal llm for recommendation systems," IEEE Access, vol. 12, no. 1, pp. 145686-145702, 2024, doi: 10.1109/ACCESS.2024.3471568.

[14] Y. Liu and J. Li, "" Expansion" and" Obstacles": The New Wave of Intelligent Media Empowering the Development of Film and Television Arts with Digital Technology," Journal of Social Science Humanities and Literature, vol. 7, no. 6, pp. 57-66, 2024, doi: 10.53469/jsshl.2024.07(06).11.

[15] Y. Zou, "The Value Logic, Challenges, and Development Strategies of Generative AI Empowering Film and Television Production," Advances in Education, Humanities and Social Science Research, vol. 15, no. 1, pp. 662-662, 2025, doi: 10.56028/ aehssr.15.1.662.2025.

[16] X. Zhu, C. Guo, H. Feng, et al., "A review of key technologies for emotion analysis using multimodal information," Cognitive Computation, vol. 16, no. 4, pp. 1504-1530, 2024, doi: 10.1007/s12559-024-10287-z.

[17] C. Yang, "Application Scenarios and Creation Paradigms of Artificial Intelligence in Digital Media Design," Computer Life, vol. 14, no. 1, pp. 50-53, 2026, doi: 10.54097/qsz2df37.

[18] C. Zhang and Q. Lei, "Intersemiotic translation of The Song of Everlasting Sorrow from narrative poetry to dance drama: a multimodal stylistics perspective," Semiotica, vol. 2026, no. 268, pp. 95-128, 2026, doi: 10.1515/sem-2024-0213.

[19] L. Zhang and J. Wang, "Hot Topics and Frontier Evolution in Music Research within the Multimodal Art Context: A Bibliometric Analysis Based on CiteSpace," Korean Science and Art Forum, vol. 44, no. 1, pp. 619-639, 2026, doi: 10.17548/ksaf.2026.01.30.619.

[20] J. Yi, Y. Tian, and Y. Zhao, "Design of red culture retrieval system based on multimodal data fusion and innovation of communication strategy path," IEEE Access, vol. 11, no. 1, pp. 134118-134125, 2023, doi: 10.1109/ACCESS.2023.3336419.

[21] Z. Su, S. Lin, L. Zhang, et al., "Multitask Learning-Based Affective Prediction for Videos of Films and TV Scenes," Applied Sciences, vol. 14, no. 11, p. 4392, 2024, doi: 10.3390/app14114391.

[22] J. Chen, K.P. Seng, J. Smith, et al., "Situation awareness in ai-based technologies and multimodal systems: Architectures, challenges and applications," IEEE Access, vol. 12, no. 1, pp. 88779-88818, 2024, doi: 10.1109/ACCESS.2024.3416370.

[23] M. Dokoupil, "Experiencing Landscape as a Visual Event: Dynamic Hyperstereoscopy and the Emergence of Perceptual Space," International Journal on Stereo & Immersive Media, vol. 9, no. 2, pp. 120-139, 2025, doi: 10.24140/ijsim.v9i2.9445.

[24] W. Yuan, J. Chen, S. Chen, et al., "Transformer in reinforcement learning for decision-making: a survey," Frontiers of Information Technology & Electronic Engineering, vol. 25, no. 6, pp. 763-790, 2024, doi: 10.1631/FITEE.2300548.

[25] S. Hu, L. Shen, Y. Zhang, et al., "On transforming reinforcement learning with transformers: The development trajectory," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 12, pp. 8580-8599, 2024, doi: 10.1109/ TPAMI.2024.3408271 .

[26] T. Chen and L. Mo, "Swin-fusion: swin-transformer with feature fusion for human action recognition," Neural Processing Letters, vol. 55, no. 8, pp. 11109-11130, 2023, doi: 10.1007/s11063-023-11367-1.

[27] L. Chen, H. Zhao, C. Shi, et al., "Enhancing multimodal perception and interaction: An augmented reality visualization system for complex decision making," Systems, vol. 12, no. 1, pp. 7-9, 2023, doi: 10.3390/systems12010007.

[28] Y. Wang, M. Guizani, and S. Hossain M, "Artificial Intelligence for Virtual Reality: State of the Art, Challenges, and Future Perspectives," ACM

Transactions on Multimedia Computing, Communications and Applications, vol. 21, no. 11, pp. 1-29, 2025, doi: 10.1145/3769090.
[29] R. Abdulrahman, A. Jamil, A. Amjad, et al., "Automated Deep Learning Approaches for Multimodal Emotion Recognition: A Review of Fusion Strategies, Modalities and Architectures," Machines and Algorithms, vol. 4, no. 3, pp. 198-214, 2025, doi: 10.66108/mna.v4i3.103.
[30] B. Liang, N. Li, G. Li, et al., "Sensing Technologies for Hand Gesture Recognition in Human-Robot Interaction: A Review," IEEE Sensors Journal, vol. 26, no. 2, pp. 1501-1519, 2025, doi: 10.1109/JSEN.2025.3635622.
[31] C. Mao, "An Intelligent Decision-Support Framework Using Mobile and Virtual Reality Technologies for Optimising Intangible Cultural Heritage Management," Decision Making: Applications in Management and Engineering, vol. 8, no. 2, pp. 881-897, 2025, doi: 10.31181/dmame8220251631.
[32] D. Wu, Z. Yang, P. Zhang, et al., "Virtual-reality interpromotion technology for metaverse: A survey," IEEE Internet of Things Journal, vol. 10, no. 18, pp. 15788-15809, 2023, doi: 10.1109/JIOT.2023.3265848.
[33] J. Ding, Y. Zhang, Y. Shang, et al., "Understanding world or predicting future? a comprehensive survey of world models," ACM Computing Surveys, vol. 58, no. 3, pp. 1-38, 2025, doi: 10.1145/3746449.