A Pre-Standardization Framework for Interoperable Cross-Modal Immersive 3D Art Environments Using NLP, Computer Vision and NeRF

Quan Zhou* and Xueying Luo

Department of Information Technology, Zhengzhou Vocational College of Finance and Taxation, Zhengzhou, Henan 450000, China
E-mail: zhouquan0027@163.com
**Corresponding Author*

Received 13 May 2026; Accepted 19 July 2026

Abstract

Interoperable immersive 3D environments require transparent interfaces that translate human artistic intent into controllable scene parameters and feedback-driven updates. This study proposes a pre-standardization framework for cross-modal 3D art environments integrating natural language processing, computer vision, conditional neural radiance field (NeRF) rendering and gaze-based feedback. A Bidirectional Encoder Representations from Transformers (BERT) model fine-tuned on ArtEmis and WikiArt encodes unstructured artistic language into semantic vectors, which are mapped to an eight-dimensional scene-control schema. Conditional NeRF renders the scene, while MediaPipe eye tracking measures visual attention and activates bounded parameter correction through cosine-alignment deviation. Experiments report a Semantic Fidelity Score (SFS) of 0.854, cross-modal cosine

Journal of ICT Standardization, Vol. 14_4, 537–562.
doi: 10.13052/jicts2245-800X.1442
© 2026 River Publishers

similarity (CCS) convergence to 0.834, a User Immersion Score (UIS) of 4.31/5.00 and a 1.8 s closed-loop response time. The framework clarifies reusable interfaces and evaluation indicators for future immersive-media standardization.

Keywords: ICT standardization, pre-standardization framework, immersive media interoperability, cross-modal semantic alignment, NeRF, eye-tracking feedback, scene-control parameter schema, WebXR, OpenXR.

1 Introduction

Immersive interactive technologies have become a relevant testbed for ICT standardization because they integrate natural language interfaces, computer vision, real-time rendering, user sensing and adaptive feedback in a single service pipeline. In virtual art creation, the key challenge is not only to generate visually attractive 3D content, but also to define a reproducible pathway through which subjective artistic intent can be translated into computable scene parameters and updated during interaction. This problem is closely related to pre-standardization research on interoperable immersive media, spatial content representation, human-centered interaction and cross-platform implementation. Accurately converting poetic or metaphorical user input into controllable 3D scene parameters is therefore important for both digital art applications and the broader standardization of multimodal ICT systems [19, 20, 28, 29].

Text-driven 3D content generation still faces several standardization-related barriers. At the semantic-input level, unstructured artistic expressions involving metaphor, emotion and sensory imagery are difficult to convert into stable machine-readable parameters [19, 20]. At the scene-control level, immersive 3D environments are sensitive to lighting, spatial density, color temperature, geometry and viewpoint-dependent rendering variables; this requires a transparent parameter schema and a traceable mapping mechanism [7, 26]. At the feedback-interface level, many systems still treat user-perception data as post hoc evaluation rather than as structured feedback that can be reused for closed-loop generation [9, 25]. These gaps limit interoperability, reproducibility and comparability across immersive content-generation systems. Existing studies provide valuable building blocks but seldom organize them into a standardization-oriented workflow. Visual-language pre-trained models and Contrastive Language–Image Pre-training (CLIP)-based methods improve image-text alignment, but their ability to represent fine-grained artistic emotion remains limited [14, 27]. Text-to-3D

methods such as DreamFusion-related pipelines and neural radiance field (NeRF)-based approaches support scene generation, but they often rely on structured prompts and do not explicitly define interoperable semantic-to-parameter interfaces [8, 11, 15, 22]. Eye-tracking and gaze-based methods can characterize user attention in virtual environments, but they are frequently used as isolated assessment tools rather than as feedback events in a cross-modal control loop [1, 2, 18]. Consequently, language parsing, scene rendering and user-perception feedback remain fragmented rather than being described as reusable ICT interfaces.

To address the semantic-mapping gap and the lack of interactive feedback, this paper proposes a pre-standardization technical framework for cross-modal immersive 3D art environments through closed-loop collaboration between natural language processing (NLP), computer vision (CV), conditional NeRF and eye-tracking feedback. In the language parsing layer, Bidirectional Encoder Representations from Transformers (BERT) fine-tuned on ArtEmis and WikiArt encodes unstructured artistic language and extracts emotional and sensory features. These features are converted into a quantifiable scene-control parameter vector through an emotion-sensory mapping layer. In the rendering layer, the parameter vector is injected into a conditional NeRF framework to drive implicit 3D scene modelling and multi-view rendering. In the feedback layer, MediaPipe eye tracking captures visual attention, and a cross-modal alignment module measures the cosine-similarity deviation between gaze-derived features and the initial semantic vector. When the deviation exceeds the preset threshold, bounded parameter correction and local re-rendering are triggered. The main contributions are: (1) a semantic-parameterization method supporting traceable conversion from artistic language to scene-control parameters; (2) a modular NLP-CV-NeRF pipeline that can serve as a candidate pre-standardization reference architecture for immersive content systems; and (3) a feedback-event interface and local differential correction mechanism that improve response efficiency while preserving semantic consistency.

2 Pre-Standardization Technical Framework for Cross-Modal 3D Artistic Environments

2.1 Semantic Interface and Scene-Parameter Mapping for Artistic Language

From a standardization viewpoint, the framework is organized around five candidate interface objects: (i) semantic input records, (ii) emotion-sensory

latent vectors, (iii) scene-control parameters, (iv) gaze-feedback events and (v) evaluation indicators. These objects do not constitute a formal standard, but they make the workflow explicit enough to support later interoperability testing and comparison across platforms.

Because unstructured artistic language contains subjective metaphors, emotional polarity and sensory imagery, it cannot be reliably transformed into physically meaningful 3D scene controls through lexical matching or fixed rules alone. From a pre-standardization perspective, the system requires a transparent semantic interface that maps free text into a stable set of scene parameters. This interface should be differentiable, reproducible and interpretable so that later systems can compare or reuse the same parameter space. The proposed semantic interface therefore defines three input-output levels: a text-input format, an emotion-sensory latent representation and an eight-dimensional scene-control vector whose components correspond to bounded visual attributes [13, 28, 29].

Suppose that the user-input artistic language text, after tokenization, yields a sequence of tokens. This sequence is fed into a domain-adaptive BERT model [10, 23] fine-tuned on the ArtEmis and WikiArt datasets. The model encodes each token through bidirectional attention under global contextual constraints. The hidden-layer output corresponding to each position is aggregated as the sentence-level semantic representation, producing a high-dimensional semantic vector.

$$\mathbf{h} = \mathrm{BERT}_{\theta}(X) \in \mathrm{R}^{d} \tag{1}$$

where θ denotes the fine-tuned BERT model parameters, X is the tokenized input sequence, $\mathbf{h}$ is the resulting semantic vector, and d is the hidden-state dimension. The vector captures the global semantic distribution of the text within an artistic and emotional context and serves as the unified semantic basis for the standardized parameter-mapping layer.

Artistic language affects 3D scene generation through two distinguishable semantic subspaces: emotional polarity and sensory imagery. The emotional subspace captures affective tone, intensity and aesthetic atmosphere, while the sensory subspace captures color warmth, spatial openness, brightness tendency and texture-related cues. To avoid manually assigning arbitrary weights, the emotion and sensory subspace weights are learned during validation by minimizing semantic reconstruction loss and scene-parameter deviation. In the final configuration, both subspaces are first projected to 256-dimensional hidden representations and then normalized through layer

normalization before concatenation [17, 24].

$$\mathbf{v}_e = \sigma(W_e\mathbf{h} + b_e) \in \mathrm{R}^{d_e} \tag{2}$$

$$\mathbf{v}_s = \sigma(W_s\mathbf{h} + b_s) \in \mathrm{R}^{d_s} \tag{3}$$

where W_e and W_s are projection matrices; b_e and b_s are bias vectors; $\sigma(\cdot)$ is the nonlinear activation function; $\mathbf{v}_e$ and $\mathbf{v}_s$ denote the emotion and sensory embeddings; and d_e and d_s are their respective dimensions. Emotional polarity, emotional intensity, color warmth, spatial density and brightness tendency together form a two-subspace semantic representation that can be further converted into standardized scene-control parameters [17, 24].

To transform the semantic representation into quantified controls that can drive scene generation, the emotional and sensory vectors are concatenated and fed into a parameter-mapping network. The network contains two fully connected hidden layers with 512 and 256 nodes, respectively, uses Gaussian Error Linear Unit (GELU) activation and dropout regularization, and outputs an eight-dimensional scene-control parameter vector. The output functions as the internal parameter schema of the proposed framework and supports later standardization of scene-control interfaces:

$$\mathbf{p} = f_\phi([\mathbf{v}_e; \mathbf{v}_s]) \in \mathrm{R}^m \tag{4}$$

where f_ϕ denotes the parameter-mapping network with parameters ϕ, m is the scene-parameter dimension, $[\mathbf{v}_e; \mathbf{v}_s]$ denotes vector concatenation, and $\mathbf{p}$ is the resulting scene-control vector. Each component of $\mathbf{p}$ corresponds to a physically interpretable and reusable scene attribute, including color temperature, spatial density, illumination intensity, contrast, texture granularity, viewpoint depth, ambient diffusion and local detail density. All components are normalized to the valid range of the corresponding physical quantity. This normalization allows the output to be transferred to the rendering layer in a consistent form and supports cross-system comparison of parameter values. The complete mapping process is shown in Figure 1.

Figure 1 illustrates how artistic language is transformed into a reusable scene-parameter schema. After tokenization and fine-tuned BERT encoding, semantic vectors are processed through a two-path decoupled structure. The emotion projection matrix and sensory projection matrix extract emotional and sensory subspace representations, which are then concatenated and fed into the parameter-mapping network to output an eight-dimensional scene-control vector. Each component corresponds to a physically interpretable attribute such as color temperature, spatial density and illumination intensity.

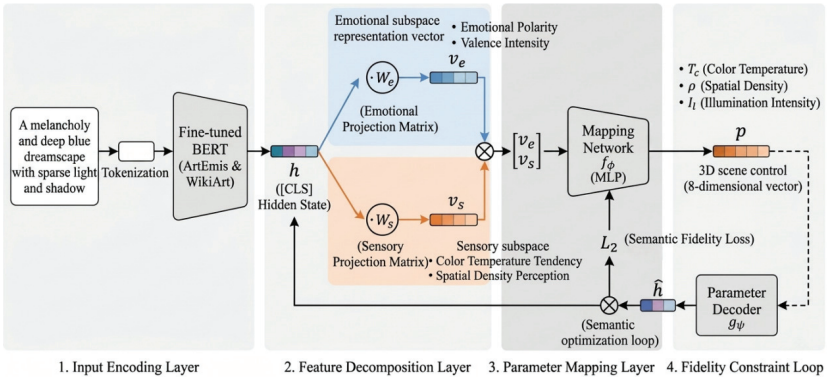


Figure 1 Structure of emotion-sensory two-branch semantic parameterization map.

The fidelity constraint loop in the lower right corner shows how a parameter decoder reconstructs the semantic vector and calculates the reconstruction deviation from the original vector. This loop helps the mapping network retain core artistic intent during parameter compression and provides a verifiable basis for semantic fidelity.

To ensure consistency between mapping results and artistic semantics, semantic fidelity is used as the training objective of the mapping network. End-to-end optimization minimizes the reconstruction deviation between the original semantic vector and the parameter-decoding vector. The reconstruction loss is defined as the squared L2 error between the original semantic vector and the decoded vector, constraining semantic compression to preserve the information needed for stable scene control:

$$\mathrm{L}_{sem} = \|\mathbf{h} - g_\psi(\mathbf{p})\|_2^2 \tag{5}$$

where g_ψ is the parameter decoder with parameters $\psi, \mathbf{h}$ is the original semantic vector, and $\mathbf{p}$ is the scene-control vector. Minimization encourages the mapping network to preserve core semantic information in the parameter space, thereby supporting traceable and verifiable transformation from artistic language to scene controls. The final parameter vector is passed as conditional input to the scene generation stage and drives NeRF-based implicit modelling of the 3D art scene.

2.2 Conditional NeRF Rendering as the Scene Realization Layer

The conditional NeRF module is therefore treated as the scene-realization layer of the candidate framework rather than as an isolated rendering

algorithm. Its role is to translate a bounded parameter schema into continuous spatial media representation, which is important for interoperability with immersive-media pipelines.

The scene-control parameter vector encodes artistic semantics in a finite-dimensional and bounded form, whereas the spatial structure of a 3D environment is continuous. Therefore, a conditional mapping mechanism is required to connect the standardized parameter schema with a continuous volume field. In the proposed implementation, the eight-dimensional parameter vector is injected into the NeRF multilayer perceptron through feature-wise affine modulation. The modulation affects density and color branches separately, enabling the same semantic-control schema to influence geometry, illumination, spatial distribution and artistic appearance while preserving a reusable interface between the scene-parameter layer and the rendering layer [11, 12, 22].

The neural radiance field model represents three-dimensional space as a continuous implicit function and outputs volume density and radiance color for each spatial location and viewing direction. Let the coordinate of the spatial sampling point be $\mathbf{r}$ and the unit vector of the viewing direction be $\mathbf{d}$. The positional encoding of the coordinate, the encoded viewing direction and the scene-control parameter vector are concatenated before entering the shared MLP trunk. In the density branch, the parameter vector modulates the probability of spatial occupancy and local density. In the color branch, it adjusts illumination, color temperature, saturation and style-related appearance at each sampling point. Inputting these quantities into the conditional radiance-field network yields the volume density and color output for that point:

$$(\sigma, \mathbf{c}) = F_{\Omega}(\gamma(\mathbf{r}), \gamma(\mathbf{d}), \mathbf{p}) \tag{6}$$

where σ denotes volume density, $\mathbf{c}$ is the RGB color vector, F_{Ω} is the conditional radiance-field network parameterized by Ω, $\gamma(\cdot)$ is the positional-encoding function, and $\mathbf{p}$ is the scene-control vector. Positional encoding maps low-dimensional coordinates to a high-frequency feature space and improves detailed structural expression. Conditional injection is implemented before the density and color heads, allowing global scene parameters to constrain both geometric distribution and view-dependent appearance. This design clarifies how the proposed parameter schema can be treated as a candidate NeRF conditional-control interface for standard-compatible immersive rendering systems [4, 6, 22].

To synthesize a two-dimensional rendered image from the implicit volume field, the radiance field is sampled along each camera ray. Let the depth

of the i-th sampling point on the ray be t_i, the corresponding volume density be σ_i, the color be $\mathbf{c}_i$, and the distance between adjacent sampling points be δ_i. The transmittance at this point is defined as:

$$T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right) \tag{7}$$

where T_i represents the cumulative probability that the ray reaches the i-th sampling point without being occluded by previous samples. It is determined by the densities σ_j and intervals δ_j of all preceding samples. The final rendered color of the corresponding pixel is obtained by weighting the color at each point according to transmittance and density.

$$\widehat{\mathbf{C}} = \sum_{i=1}^{N} T_i (1 - \exp(-\sigma_i \delta_i)) \mathbf{c}_i \tag{8}$$

where N denotes the total number of sampling points on a ray, and $\mathbf{C}$ denotes the rendered pixel color; T_i, σ_i, δ_i and $\mathbf{c}_i$ retain the meanings defined above. This integration process is performed over all pixel rays to complete volumetric rendering of the full-view image.

As shown in Figure 2, the volume-rendering process of the parameterized neural radiance field connects the semantic-parameter space and the visual-image space into a complete scene-realization path.

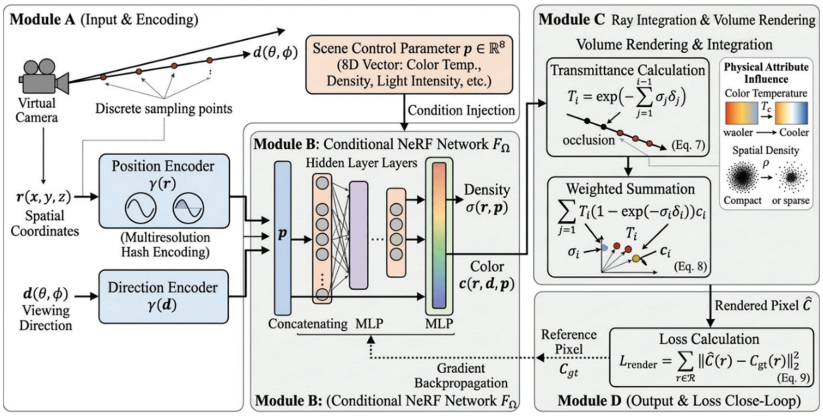


Figure 2 Schematic diagram of the parametrically conditional neural radiance field volume rendering process.

Spatial coordinates and viewing directions are mapped to a high-frequency feature space by a position encoder and then input into a conditional radiance-field network together with the scene-control parameter vector. This conditional injection imposes global constraints on the mapping behavior of the network, ensuring that volume density and color output are semantically driven by the parameter schema. Transmittance is progressively accumulated at each sampling point along the ray, and the product of volume density and color is weighted and integrated to synthesize pixel colors. The reconstruction error between rendered and reference pixels drives network convergence through gradient backpropagation, forming a closed optimization loop for output and loss. This process reveals how artistic semantic parameters are transmitted from conditional injection and implicit volume-field inference to image synthesis.

The network parameters are optimized by minimizing the reconstruction error between the rendered color and the reference image. The training objective is defined as:

$$\mathrm{L}_{render} = \sum_{\mathbf{r}\in\mathrm{R}} \|\widehat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}_{gt}(\mathbf{r})\|_2^2 \tag{9}$$

where R denotes the sampled ray set in the current batch, **C(r)** is the rendered pixel color for ray $\mathbf{r}$, and $\mathbf{C}_{gt}(\mathbf{r})$ is the corresponding reference pixel color. The optimization process allows the radiance-field network to fit the geometric structure and appearance distribution of the target scene under parameter constraints, ensuring that the generated scene visually aligns with the intent encoded by the semantic parameters. After optimization, the fixed parameters support real-time inference from arbitrary viewpoints and output a multi-view rendered sequence of the initial 3D art scene. This rendered result serves as visual input for the feedback layer and participates in cross-modal semantic alignment together with eye-tracking attention features.

2.3 Visual Attention Feedback Interface and Closed-Loop Parameter Correction

For closed-loop interaction, the feedback channel is defined as a structured event interface rather than a raw sensor stream. Each event record contains a timestamp, confidence score, normalized heatmap region, Cross-modal Cosine Similarity (CCS) value, triggered parameter index and update magnitude. This design improves traceability and reduces dependence on device-specific gaze formats.

The initial rendered scene is generated from static parameter vectors and cannot by itself determine whether the user attends to scene regions that match the original artistic intent. For an interactive immersive system, the feedback interface should convert user-perception signals into measurable semantic deviation and then into bounded parameter correction. This establishes a closed-loop path from gaze sensing to alignment evaluation and rendering update. The feedback data are treated as event-level interaction records rather than biometric identity data, which is important for privacy-aware interface design [5, 16, 18].

The MediaPipe eye-tracking module acquires the user's binocular gaze-coordinate sequence at frame-level frequency. Raw gaze points are first filtered by removing frames with low tracking confidence, then normalized to screen coordinates and temporally aggregated within a sliding window of 15 frames. A Gaussian kernel density estimator converts the normalized gaze points into a visual-attention heatmap. Short-term jitter is reduced through exponential smoothing, and missing frames are handled by interpolation within the same interaction window. H and W represent the image height and width, respectively [3, 16]. To transform the two-dimensional attention distribution into a feature representation with the same dimension as the semantic vector, a lightweight attention encoder is introduced:

$$\mathbf{z} = E_{\phi}(\mathbf{A}) \in \mathrm{R}^{d} \tag{10}$$

where $\mathbf{A}$ is the visual-attention heatmap, E_{ϕ} is the attention encoder parameterized by $\phi, \mathbf{z}$ is the gaze-derived embedding, and d is the embedding dimension shared with the semantic vector. The attention encoder contains three convolutional blocks with 3×3 kernels, channel sizes of 32, 64 and 128, followed by global average pooling and a fully connected projection layer. The output is linearly projected to the same dimension as the BERT-derived semantic vector. This dimensionality-matching strategy allows the gaze-derived attention feature and the text-derived semantic feature to be compared in a shared embedding space.

To quantify the semantic discrepancy between real-time visual perception and the original artistic intent, cosine similarity is calculated in the shared embedding space [5]:

$$S = \frac{\mathbf{z} \cdot \mathbf{h}}{\|\mathbf{z}\| \cdot \|\mathbf{h}\|} \tag{11}$$

where S denotes the CCS value, while $\mathbf{z}$ and $\mathbf{h}$ are the gaze-derived and text-derived semantic embeddings, respectively. A lower value indicates a larger

deviation between the current scene presentation and the intended semantics. A deviation trigger threshold is set. When the threshold is reached, the system determines that semantic alignment has failed and activates the parameter-difference correction process. In practical deployment, gaze sequences should be processed locally whenever possible, stored only as anonymized heatmap-level records and used after informed consent, because eye-tracking data may reveal attention habits and sensitive user preferences [16, 18].

After a deviation is triggered, the semantic alignment error is converted into a correction value in the scene-parameter space. The semantic deviation vector is defined as follows, then fed into the difference-mapping network to output the parameter correction increment:

$$\Delta\mathbf{p} = f_{\phi}^{\Delta}(\Delta\mathbf{h}) \in \mathrm{R}^{m} \tag{12}$$

where $\Delta\mathbf{p}$ is the scene-parameter correction vector, f_{ϕ}^{Δ} is the difference-mapping network with independent parameters $\overset{\Delta}{\phi}$, $\Delta\mathbf{h}$ denotes the semantic-deviation vector in the shared embedding space, and m is the scene-parameter dimension. The network shares the structural design of the parameter-mapping network in Section 2.1 and outputs a correction vector with the same scene-parameter dimension, so each component corresponds to the adjustment magnitude of attributes such as color temperature, spatial density and illumination intensity. The correction is not a free global regeneration step. Instead, only parameter components whose normalized deviation exceeds the local update threshold are updated, which avoids unnecessary disturbance to semantically stable regions.

The corrected scene parameters are obtained by superimposing the original parameter vector with the correction increment:

$$\mathbf{p}' = \mathbf{p}^{+\lambda}\Delta\mathbf{p} \tag{13}$$

where $\mathbf{p}'$ is the updated scene-control vector, $\mathbf{p}$ is the original vector, $\Delta\mathbf{p}$ is the correction increment, and λ is the step-size coefficient controlling the magnitude of a single update and prevents overshoot or semantic oscillation. The update follows cosine-alignment optimization rather than reinforcement learning. Iteration stops when the CCS value exceeds 0.82 for two consecutive frames, when the change in scene-control parameters falls below 0.01, or when the maximum of 10 correction rounds is reached. The updated conditional input is re-injected into the radiance-field network, triggering local voxel resampling and rendering refresh in the region covered by the attention heatmap. Rendering caches in non-deviation regions remain unchanged, balancing correction accuracy and computational overhead [6, 21].

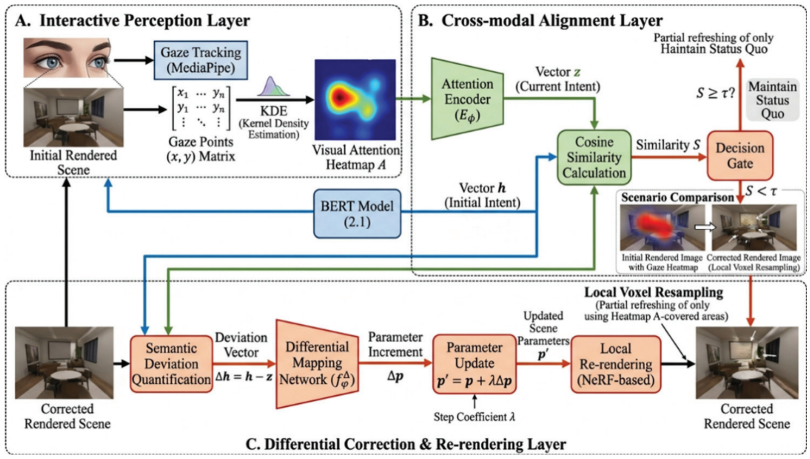


Figure 3 Architecture diagram of eye-tracking driven cross-modal closed-loop feedback and parameter difference correction.

As shown in Figure 3, the closed-loop feedback process consists of three levels: perception, alignment and correction. The MediaPipe module extracts the gaze sequence, performs confidence filtering and temporal aggregation, and generates an attention heatmap through Gaussian kernel density estimation. The attention encoder maps the heatmap into an embedding vector, which is compared with the initial BERT-derived semantic vector through cosine similarity. The threshold node acts as the gating mechanism. When deviation is detected, the semantic deviation vector is transformed into parameter increments through the difference-mapping network and superimposed on the original parameters under the step-size constraint. The updated scene parameters are re-injected into the NeRF radiance field, and local voxel resampling and rendering refresh are performed only in the deviated heatmap regions. The corrected scene image then re-enters the eye-tracking acquisition process for the next frame, allowing the system to approach the user's intent iteratively.

3 Experimental Setup and Verification Design

3.1 Experimental Environment and Reproducible Implementation Conditions

The system runs on a workstation platform equipped with an NVIDIA RTX 3090 GPU with 24 GB of VRAM, an Intel Core i9-12900K CPU and 64

GB of DDR5 RAM. The operating system is Ubuntu 20.04 LTS, the deep learning framework is PyTorch 1.13.1 with CUDA 11.7, and the NeRF rendering framework is implemented based on instant-ngp to support real-time volumetric inference accelerated by hash encoding. The MediaPipe eye-tracking module is deployed on the same workstation and captures gaze-coordinate sequences through an RGB camera at 30 fps with approximately 0.5 degrees of visual-angle error. The implementation is reported to improve reproducibility and provide a reference configuration for later comparative evaluation across OpenXR/WebXR-compatible immersive clients and MPEG-I-oriented spatial media workflows.

The BERT model uses BERT-base-uncased pre-trained weights and is fine-tuned on the ArtEmis and WikiArt datasets for domain-adaptive optimization. The ArtEmis dataset contains over 81,000 sentiment-annotated artwork records, while the WikiArt dataset covers 81,444 artworks across 27 art styles with corresponding descriptive texts. The merged dataset is divided into training, validation and test sets in an 8:1:1 ratio. During fine-tuning, the learning rate is 2×10^{-5}, the batch size is 32, the number of epochs is 10, dropout is 0.1 and weight decay is 0.01. The AdamW optimizer is used with linear learning-rate warm-up. The loss function combines sentence-level semantic reconstruction loss, emotion-label cross-entropy and parameter-decoder reconstruction loss, ensuring that the language encoder supports both artistic understanding and scene-control mapping.

During NeRF network training, the number of sampling points per ray is set to 128, the scene bounding box is normalized to a unit cube and the position-encoding frequency order is set to 10. The differential mapping network and the parameter-mapping network are structurally aligned. The correction step-size coefficient is set to 0.3, the cross-modal alignment deviation threshold is set to 0.75 and the Gaussian kernel bandwidth used for attention-heatmap generation is adaptively set according to screen resolution. During closed-loop iteration, step size is reduced by 20% if the CCS value decreases after an update. Local update regions are selected from the top 20% of heatmap intensity values, while unchanged regions reuse cached volume-rendering results. The key hyperparameters are summarized in Table 1.

The parameters in Table 1 are fixed during system deployment and are not dynamically adjusted with user input. This setting ensures comparability and reproducibility under different artistic-language input conditions and supports the use of the results as an experimental reference for future standardization-oriented studies.

Table 1 System key hyperparameter configuration

Parameter	Symbol	Value	Module
Affective subspace dimension	d_e	256	Semantic mapping layer
Sensory subspace dimension	d_s	256	Semantic mapping layer
Scene parameter dimension	m	8	Parameter mapping network
Ray sampling points	N	128	NeRF rendering
Correction step size	λ	0.3	Differential correction network
Deviation trigger threshold	τ	0.75	Cross-modal alignment module
BERT fine-tuning learning rate	–	2×10^{-5}	Language parsing module

3.2 Evaluation Indicators and Comparative Protocol

The evaluation design was further aligned with standardization-oriented assessment. Instead of reporting only algorithmic accuracy, the study reports semantic, interaction and response indicators that can be reused by future immersive-system benchmarks. The evaluation protocol is constructed around four dimensions relevant to immersive ICT systems: artistic semantic fidelity, cross-modal alignment quality, user immersion and system response efficiency. The Semantic Fidelity Score (SFS), termed Artistic Semantic Fidelity (ASF) in the cross-method benchmark in Table 2, is calculated as the mean cosine similarity between the decoder-reconstructed semantic vector and the original text-derived semantic vector. CCS measures the alignment between the eye-tracking semantic feature vector and the initial artistic semantic vector according to Equation (11). User Immersion Score (UIS) is based on a five-point Likert scale covering spatial immersion, artistic semantic fit, visual comfort, interaction naturalness, responsiveness and overall satisfaction. Scene Generation Time (SGT) records the time from deviation triggering to completion of local re-rendering. Together, these indicators form a candidate benchmarking set for comparing interoperable immersive-generation systems.

The comparative design examines both the closed-loop feedback mechanism and the overall method performance. At the module-ablation level, three configurations are compared: a pure-text static baseline that retains language parsing and static NeRF generation without visual feedback; a CV-augmented non-feedback method that incorporates computer-vision features without parameter correction; and the complete NLP-CV closed-loop method proposed in this paper. To strengthen baseline coverage, the comparison also distinguishes text-driven generation pipelines, non-closed-loop cross-modal alignment pipelines and editing-oriented NeRF pipelines. All configurations run on the same hardware platform and use the same hyperparameters. At

the method-comparison level, seven representative methods, DreamFusion, Magic3D, ProlificDreamer, CLIP-NeRF, LERF, NeRF-Art and Instruct-NeRF2NeRF, are selected as benchmarks. They cover pure text-driven generation, cross-modal semantic alignment and artistic scene editing. All benchmark methods are evaluated on a unified test set using SFS/ASF, CCS, UIS and SGT. The user experiment recruits participants with art-related aesthetic experience.

Participants freely input unstructured artistic language and complete the immersive scene experience without guidance, and scoring data are collected immediately after the experience.

4 Results and Discussion

4.1 Artistic Semantic Fidelity of the Language-Parameter Interface

To verify how different encoding strategies affect semantic fidelity in the language-parsing module, six schemes are compared under five categories of unstructured artistic language input: figurative description, emotional color, spatial imagery, abstract metaphor and cross-sensory synesthesia. The six schemes are BERT without fine-tuning, ArtBERT for art-emotion pre-training, CLIP text encoder, BERT fine-tuned on ArtEmis, BERT fine-tuned on WikiArt and the dual-dataset joint fine-tuning BERT scheme adopted in this paper. All methods are run under the same samples and evaluation protocol. The results are shown in Figure 4.

As shown in Figure 4, the semantic fidelity score of the proposed method reaches an average of 0.854 across five language styles, an improvement of 0.249 over the baseline BERT score of 0.605. It also maintains a clear advantage over ArtBERT (0.700) and the CLIP text encoder (0.691). Scores decrease as input abstraction increases, indicating that metaphorical language places higher demands on the semantic interface. The baseline BERT score falls to 0.492 for cross-sensory synesthetic inputs, with a standard deviation of 0.058, showing severe degradation in subjective artistic-language processing. The ArtEmis-tuned BERT performs better on emotionally charged inputs, while the WikiArt-tuned BERT performs better on spatial imagery inputs, confirming the complementarity of emotional and visual-spatial datasets. The proposed joint fine-tuning method has the lowest standard deviations for abstract metaphor and cross-sensory synesthesia, indicating stronger robustness of the semantic-to-parameter interface.

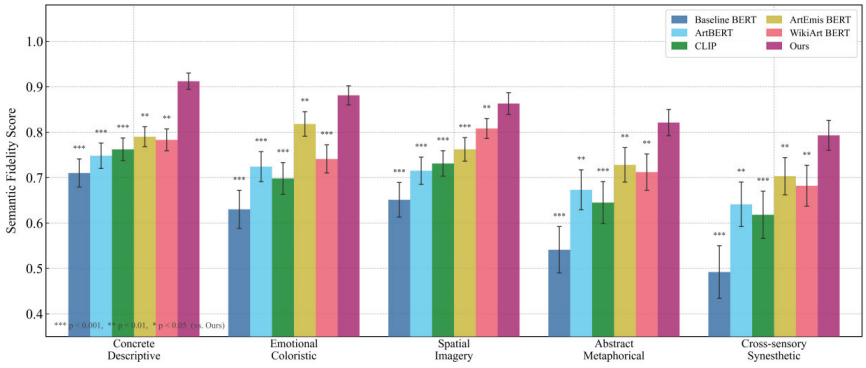


Figure 4 Comparison of semantic fidelity scores of different language encoding schemes under five artistic language styles.

Note: *** indicates $p < 0.001$, ** indicates $p < 0.01$, * indicates $p < 0.05$. Significance was assessed using independent-samples t-tests comparing each method with the proposed BERT model jointly fine-tuned on the two datasets.

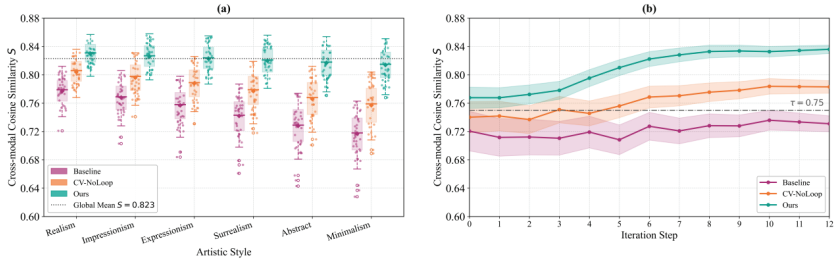


Figure 5 Comparison of cross-modal semantic alignment performance under different art styles and iteration processes: (a) cross-modal cosine similarity distributions for three methods across six art styles; (b) convergence evolution curves of cross-modal cosine similarity during closed-loop iteration.

4.2 Cross-Modal Semantic Alignment and Feedback Stability

To evaluate alignment quality and dynamic correction under different artistic-language conditions, cosine similarity is used as the core metric. Under unstructured artistic text inputs representing six styles (Realism, Impressionism, Expressionism, Surrealism, Abstract Art, Minimalism), three methods are compared: a text-only static baseline, a CV-augmented non-feedback method and the proposed NLP-CV closed-loop collaborative method. The styles are arranged from lower to higher metaphor density to examine alignment stability and iterative convergence in increasingly unstructured artistic semantic spaces. The results are shown in Figure 5.

As shown in Figure 5(a), as metaphor density increases, the median similarity of the baseline declines from 0.779 for Realism to 0.718 for Minimalism, with a wider interquartile range. Low-value outliers appear in Surrealism, Abstract Art and Minimalism, showing systematic degradation of a purely static text-mapping path. The CV-NoLoop method improves similarity across all styles but still remains below the global mean in Abstract Art and Minimalism, indicating that visual features alone are insufficient without feedback correction. In contrast, the NLP-CV-Loop method maintains medians between 0.815 and 0.831 across styles, with a stable interquartile range and few outliers. As shown in Figure 5(b), the average similarity of NLP-CV-Loop increases after crossing the deviation threshold in steps 4 to 6 and converges to 0.8337 by step 9, while the confidence interval narrows. These results show that eye-tracking feedback and parameter-difference correction jointly improve alignment stability and convergence efficiency.

4.3 Closed-Loop Iterative Convergence and Rendering Efficiency

To verify the efficiency of parameter-difference correction and local re-rendering, eight typical art styles are selected as input conditions: Impressionism, Surrealism, Baroque, Minimalism, Abstract Expressionism, Pop Art, Realism and Cubism. Under the same hardware and hyperparameter settings, local and global re-rendering strategies are compared. Single-iteration convergence time is recorded for each style, and cross-modal semantic alignment cosine similarity is used to track scene-quality evolution during 10 iterations. The results are shown in Figure 6.

As shown in Figure 6, the local re-rendering strategy has lower single-round convergence time than global re-rendering across all eight art styles. Local re-rendering ranges from 1.4 s to 2.3 s, whereas global re-rendering ranges from 3.6 s to 7.2 s, saving about 62% of the time on average. Scene spatial complexity amplifies the difference: in Minimalism the difference is 2.2 s, while in Realism it expands to 4.9 s. This indicates that local re-rendering, which updates only the deviation region, is particularly advantageous in complex scenes. In iterative quality evolution, local re-rendering crosses the threshold $\tau = 0.75$ in the third round and converges around 0.823, whereas global re-rendering crosses the threshold in the fifth round and approaches the same value by the tenth round. Thus, local re-rendering reduces response time without lowering the semantic-alignment ceiling, which is important for real-time interactive system implementation.

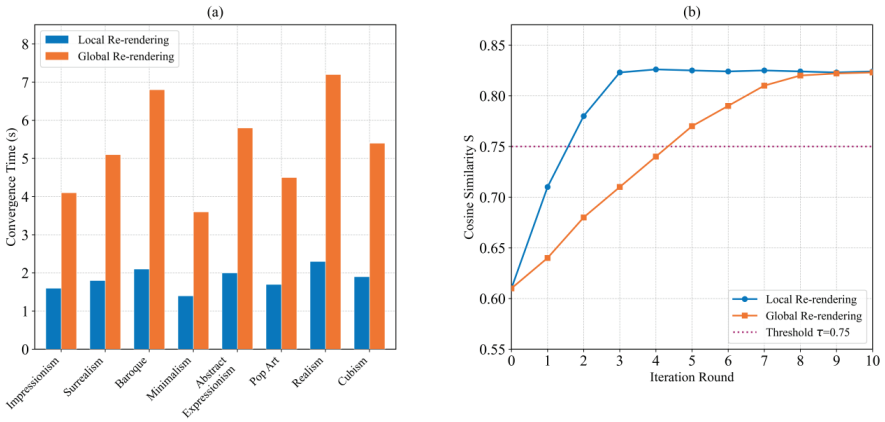


Figure 6 Experimental results comparing local re-rendering and global re-rendering strategies: (a) comparison of single-round iteration convergence time of two rendering strategies under eight art styles; (b) cross-modal semantic alignment cosine similarity change curves of the two rendering strategies during 10 iterations.

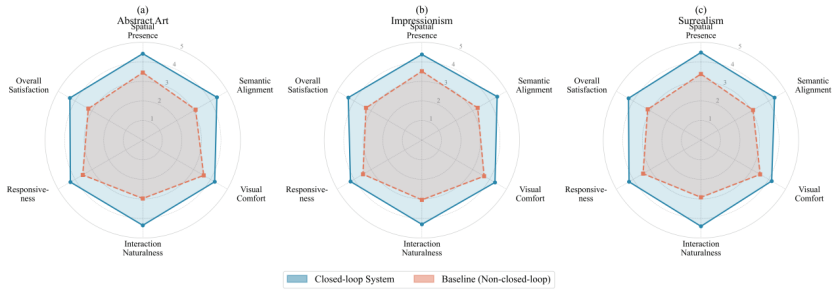


Figure 7 Comparison of six-dimensional User Immersion Scores for closed-loop and non-closed-loop systems under different art styles: (a) Abstract Art, (b) Impressionist, (c) Surrealist.

4.4 User Immersion and Perceived Interaction Quality

To verify whether the closed-loop mechanism improves user immersion, subjective rating data are collected under two configurations: the closed-loop system and a non-closed-loop baseline system. The evaluation dimensions include spatial immersion, artistic semantic fit, visual comfort, interaction naturalness, responsiveness and overall satisfaction. Each dimension has a maximum score of 5.00, and participants complete a Likert-scale assessment after interaction. Figure 7 compares six-dimensional user immersion under three art styles.

The six-dimensional mean scores of the closed-loop system under Abstract Art, Impressionism and Surrealism are 4.33, 4.34 and 4.32, respectively, while the non-closed-loop baseline scores are 3.32, 3.39 and 3.24. The closed-loop system therefore shows a clear advantage across all styles. The largest inter-group difference appears in interaction naturalness under Surrealism, reaching 1.48 points. This result is consistent with the highly metaphorical and unstructured nature of Surrealist language. Without eye-tracking feedback correction, the baseline system struggles to capture deeper intent, resulting in lower perceived interaction quality. Under Impressionism, the closed-loop score for artistic semantic fit reaches 4.45, while responsiveness is 4.22, suggesting that richer visual detail can increase rendering pressure. Visual comfort remains relatively stable under non-closed-loop conditions, indicating that basic visual quality is mainly supported by the NeRF rendering layer. Overall, the closed-loop mechanism provides the greatest benefit when semantic complexity is high.

4.5 Overall System Performance Against Benchmark Methods

To evaluate the comprehensive performance of the proposed framework across semantic understanding, cross-modal alignment, user experience and response efficiency, it is compared with seven representative methods. These include pure text-driven 3D generation methods such as DreamFusion, Magic3D and ProlificDreamer; cross-modal semantic alignment methods such as CLIP-NeRF and LERF; and art-oriented or interactive editing methods such as NeRF-Art and InstructNeRF2NeRF. All methods are run on a unified hardware platform and the same unstructured artistic-language test set. The metrics are ASF, CCS, UIS and SGT. The comparison results are shown in Table 2.

As shown in Table 2, performance follows a gradient that is consistent with module completeness. The three pure text-driven generation methods have ASF scores from 0.614 to 0.661, and CCS values below 0.620, mainly because they are designed for structured prompts and lack an explicit semantic-to-parameter interface for unstructured artistic language. CLIP-NeRF and LERF improve CCS through language-visual alignment, but without closed-loop correction their ASF remains below 0.710 and UIS stays between 3.35 and 3.41. NeRF-Art improves artistic style perception, but it cannot dynamically correct deviations from user intent. InstructNeRF2NeRF achieves the best benchmark performance, with ASF and CCS of 0.768 and 0.751, but its reliance on structured instructions limits its ability to handle

Table 2 Comparison of the overall performance of the proposed method with existing benchmark methods

Method	ASF↑	CCS↑	UIS ↑ (/5.00)	SGT↓(s)
DreamFusion	$0.614 \pm 0.032^{***}$	$0.571 \pm 0.028^{***}$	$2.87 \pm 0.31^{***}$	$8.6 \pm 0.74^{***}$
Magic3D	$0.638 \pm 0.029^{***}$	$0.589 \pm 0.031^{***}$	$3.02 \pm 0.27^{***}$	$7.2 \pm 0.68^{***}$
ProlificDreamer	$0.661 \pm 0.026^{***}$	$0.612 \pm 0.024^{***}$	$3.18 \pm 0.29^{***}$	$9.4 \pm 0.81^{***}$
CLIP-NeRF	$0.693 \pm 0.023^{***}$	$0.674 \pm 0.021^{***}$	$3.35 \pm 0.24^{***}$	$5.3 \pm 0.57^{***}$
LERF	$0.701 \pm 0.021^{***}$	$0.689 \pm 0.019^{***}$	$3.41 \pm 0.22^{***}$	$4.8 \pm 0.52^{***}$
NeRF-Art	$0.724 \pm 0.019^{***}$	$0.703 \pm 0.022^{***}$	$3.57 \pm 0.21^{***}$	$4.1 \pm 0.49^{***}$
InstructNeRF2NeRF	$0.768 \pm 0.017^{**}$	$0.751 \pm 0.018^{**}$	$3.89 \pm 0.19^{**}$	$3.2 \pm 0.43^{**}$
Ours	$\mathbf{0.847 \pm 0.014}$	$\mathbf{0.823 \pm 0.015}$	$\mathbf{4.31 \pm 0.17}$	1.8 ± 0.38

Note: $^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$. Significance was assessed relative to the proposed method using independent-samples t-tests for the performance differences between each benchmark and the proposed method.

metaphorical input, and its SGT of 3.2 s leaves a gap in response efficiency. The proposed method achieves ASF of 0.847, CCS of 0.823, UIS of 4.31/5.00 and SGT of 1.8 s, with the narrowest error range. These results validate the overall effectiveness of the NLP-CV closed-loop framework and its relevance as a technical reference for emerging immersive-content standardization. The benchmark results should be interpreted as evidence of the effectiveness of the proposed modular workflow under the reported experimental conditions, rather than as a claim that the framework is a completed standard. This distinction keeps the contribution within the scope of pre-standardization research.

4.6 Standardization Implications, Interoperability Interfaces and Implementation Considerations

Although the proposed system is not presented as a formal ICT standard, the results identify several standardization-relevant interfaces that can support future work by standards development organizations, consortia or open-source communities. The first is the semantic input interface, where unstructured artistic language is encoded into a reproducible vector representation. The second is the scene-control parameter schema, which translates semantic features into bounded and physically interpretable controls. The third is the NeRF conditional-control interface, where scene parameters are injected into density and color branches through a stable modulation path. The fourth is the feedback-event interface, where gaze heatmaps are converted into semantic-deviation signals. The fifth is the evaluation interface, where SFS/ASF, CCS,

UIS and SGT provide comparable indicators for semantic fidelity, alignment, user experience and response efficiency.

These interfaces are useful for pre-standardization because they separate the system into modular layers that can be implemented, replaced or benchmarked independently. The design is compatible with the direction of OpenXR and WebXR in cross-platform interaction access, ISO/IEC 23090 and MPEG-I in immersive media representation, and XR safety or user-comfort concerns such as those reflected in UL 8400. For example, different language encoders could share the same scene-parameter schema, and different rendering engines could be evaluated using the same feedback and alignment metrics. The closed-loop design also highlights governance issues that future ICT standards may need to address, including transparency of semantic mapping, reproducibility of user-feedback processing, privacy protection for gaze data and interoperability between immersive-content authoring tools. Therefore, the proposed framework contributes not only a generation method but also a structured reference architecture for standardizable cross-modal immersive media workflows [28, 29].

5 Conclusion and Standardization Outlook

This study addresses two linked problems: the semantic-mapping gap between unstructured artistic language and immersive 3D space, and the absence of user-perceptual feedback in static generation systems. It constructs a closed-loop collaborative framework based on NLP, CV, conditional NeRF and eye-tracking feedback. Experiments demonstrate advantages in semantic fidelity, cross-modal alignment, user immersion and response efficiency, validating the effectiveness of emotion-sensory parameter mapping, conditional NeRF modelling and attention-driven differential correction. From the Journal of ICT Standardization perspective, the value of the work lies not in proposing a formal standard, but in defining reusable interfaces, measurable indicators and an interoperable workflow for emerging immersive 3D content systems. Current limitations include cultural-style bias in fine-tuned datasets, restricted generalization to cross-cultural metaphors and manually set hyperparameters that may not adapt to individual users. Future work will expand multicultural corpora, introduce dynamic threshold strategies, extend feedback signals to physiological modalities such as EEG and further refine the framework into a clearer reference model for pre-standardization research in cross-modal immersive ICT systems.

Acknowledgements

This work was supported by the following projects: the Henan Provincial Federation of Social Sciences Research Project "Applied Research on Cultural and Creative Product Design Based on Henan Red Culture" (August 2022, Participant); the Zhengzhou Social Science Research Project "Research on Innovative Pathways for Ideological and Political Education in Zhengzhou Universities in the Era of Artificial Intelligence" (May 2025, Participant); and the Institutional Teaching Reform Project of Zhengzhou Vocational College of Finance and Taxation "Research on Innovation in Computer Teaching in Higher Vocational Education Empowered by Generative Artificial Intelligence" (June 2025, Participant).

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Choi, H., and Nam, S. (2025). Effect of eye-tracking-based depth perception visualization interface on virtual reality user experience. Virtual Reality, 30(1), 5.

[2] Choi, H., Kwon, J., and Nam, S. (2023). Research on the application of gaze visualization interface on virtual reality training systems. Journal on Multimodal User Interfaces, 17(3), 203–211.

[3] Fehlinger, P., Ertl, B., and Watzka, B. (2026). Eye movement analysis: A kernel density estimation approach for saccade direction and amplitude. Journal of Eye Movement Research, 19(1), 10.

[4] Gao, C., Sun, Q., Zhu, J., and Chen, J. (2025). MBS-NeRF: Reconstruction of sharp neural radiance fields from motion-blurred sparse images. Scientific Reports, 15(1), 5275.

[5] Ghamati, K., Dehkordi, M. B., and Zaraki, A. (2025). Which AI sees like us? Investigating the cognitive plausibility of language and vision models via eye-tracking in human-robot interaction. Sensors, 25(15), 4687.

[6] Gu, X., Chang, Y., Xia, J., Yu, Y., Tian, Z., and Chen, J. (2026). OptiNeRF: A spatially optimized neural rendering framework for complex scene reconstruction. Mathematics, 14(5), 842.

[7] Gu, Z., Gao, T., and Liu, H. (2025). Text-to-3D scene generation framework: Bridging textual descriptions to high-fidelity 3D scenes. Visual Computing for Industry, Biomedicine, and Art, 8(1), 29.

[8] He, Y., Bai, Y., Lin, M., Sheng, J., Hu, Y., Wang, Q., … and Liu, Y. J. (2024). Text-image conditioned diffusion for consistent text-to-3D generation. Computer Aided Geometric Design, 111, 102292.

[9] Huang, Y. T., Hsu, C. C., and Wang, T. H. (2025). Effects of interactive loading interfaces for virtual reality game environments on time perception, cognitive load, and emotions. Frontiers in Virtual Reality, 6, 1540406.

[10] Ishikawa, S., and Sugiura, K. (2023). Affective image captioning for visual artworks using emotion-based cross-attention mechanisms. IEEE Access, 11, 24527–24534.

[11] Li, H., Tian, Y., Lan, K., Liao, Y., Wang, L., Hui, P., and Zhou, P. Y. (2025). DreamScene: 3D Gaussian-based end-to-end text-to-3D scene generation. IEEE Transactions on Pattern Analysis and Machine Intelligence. doi: 10.1109/TPAMI.2025.3620403.

[12] Ma, R., and He, S. (2025). Multi-channel volume density neural radiance field for hyperspectral imaging. Scientific Reports, 15(1), 16253.

[13] Ning, X., Yu, Z., Li, L., Li, W., and Tiwari, P. (2024). DILF: Differentiable rendering-based multi-view image-language fusion for zero-shot 3D shape understanding. Information Fusion, 102, 102033.

[14] Pei, E., Zhao, H., Zhang, T., Jiang, D., He, L., and Chen, H. (2025). Multi-modal prompt learning for facial expression recognition: Leveraging emojis and large language models. Information Fusion, 104063.

[15] Peng, X., Sun, T., Hu, Q., Sun, Z., Xu, N., and Peng, J. (2025). Poe2CLP: Phrase-level attention and cross-modal semantic alignment for poem-generate Chinese landscape paintings. npj Heritage Science, 13(1), 656.

[16] Rodrigues, J. A., Castro, A. V. D., and Llamas-Nistal, M. (2026). Advanced consumer behaviour analysis: Integrating eye tracking, machine learning, and facial recognition. Journal of Eye Movement Research, 19(1), 9.

[17] Stamkou, E., Keltner, D., Corona, R., Aksoy, E., and Cowen, A. S. (2024). Emotional palette: A computational mapping of aesthetic experiences evoked by visual art. Scientific Reports, 14(1), 19932.

[18] Stark, P., Hasenbein, L., Kasneci, E., and Goellner, R. (2024). Gaze-based attention network analysis in a virtual reality classroom. MethodsX, 12, 102662.

[19] Su, C., Wang, X., Liu, S., and Chen, Y. (2024). Efficient visual metaphor image generation based on metaphor understanding. Neural Processing Letters, 56(3), 150.

[20] Xu, C., and Zhou, S. (2025). Visualizing poetry with deep semantic understanding and consistency evaluation. npj Heritage Science, 13(1), 686.

[21] Ye, C., Chen, X., Chen, Z., Sun, Z., Wu, S., and Deng, W. (2025). UAVEdit-NeRFDiff: Controllable region editing for large-scale UAV scenes using neural radiance fields and diffusion models. Symmetry, 17(12), 2069.

[22] Zhang, J., Li, X., Wan, Z., Wang, C., and Liao, J. (2024). Text2NeRF: Text-driven 3D scene generation with neural radiance fields. IEEE Transactions on Visualization and Computer Graphics, 30(12), 7749–7762.

[23] Zhang, M., Hou, X., Yan, Y., and Sun, M. (2024). DecoupleCLIP: A novel cross-modality decouple model for painting captioning. Electronics, 13(21), 4207.

[24] Zhao, Q., Xia, Y., Long, Y., Xu, G., and Wang, J. (2025). Leveraging sensory knowledge into text-to-text transfer transformer for enhanced emotion analysis. Information Processing & Management, 62(1), 103876.

[25] Zhao, Y., and Cen, Z. (2024). Exploring multimodal feedback mechanisms for improving user interaction in virtual reality environments. Journal of Industrial Engineering and Applied Science, 2(6), 35–41.

[26] Zhou, Y., He, Z., Li, Q., and Wang, C. (2026). LayoutDreamer: Physics-guided layout for text-to-3D compositional scene generation. Pattern Recognition, 113427.

[27] Zhou, Z., Zhai, Z., Gao, X., and Zhu, J. (2025). Improved IEC performance via emotional stimuli-aware captioning. Scientific Reports, 15(1), 22173.

[28] Liu, Y. (2025). Data analytics in the Internet of Things era: Tools, approaches, challenges, and solutions. Journal of ICT Standardization, 13(4), 427–448.

[29] Lalli, G. (2025). Defining interoperability: A universal standard. Journal of ICT Standardization, 13(2), 139–156.

Biographies



Quan Zhou is a lecturer at Zhengzhou Vocational College of Finance and Taxation. She holds a master's degree and is from Zhengzhou, Henan Province, China. Her research interests focus on digital media art, animation education, and the integration of regional culture with digital artistic expression. She is engaged in teaching and research in the field of animation at the vocational education level. In recent years, she has been dedicated to curriculum reform and pedagogical innovation for animation programs, as well as exploring new approaches to the application of traditional regional culture in digital art education. She has led and participated in several provincial and municipal-level educational research projects, published multiple academic papers, and authored two books and textbooks. She has also been granted one utility model patent. Her honors include the Second Prize in a Provincial Teaching Ethics and Professional Conduct Essay Competition and the Second Prize in a College-level Teaching Competition.



Xueying Luo is an assistant lecturer at Zhengzhou Vocational College of Finance and Taxation. She holds a master's degree and is from Weinan, Shaanxi Province, China. Her research interests include computer application technologies and artificial intelligence applications. She is currently engaged

in teaching and research activities related to artificial intelligence education. Her recent work focuses on teaching reform and curriculum development for artificial intelligence programs in higher vocational education. She has published several academic papers in related fields and is committed to promoting the integration of emerging technologies with vocational education.