KG-GAN: Knowledge Graph-Constrained GAN for Culturally Faithful Virtual Scene Synthesis

Peishu Song
Personnel division,Jilin Polytechnic of Water Resources and Electric Engineering, Changchu,130117, China
E-mail:songpeishu1991@outlook.com

Keywords: virtual cultural scene, knowledge-guided generative adversarial network, knowledge graph constraints, semantic consistency/fidelity, immersive experience optimization, temporal stability, attention mechanism

Received: October 4, 2025

This paper presents a knowledge-guided generative adversarial network (KG-GAN) for culturally faithful virtual scene construction and immersive experience optimization. A domain knowledge graph (KG) encodes traditional cultural entities and relations as graph embeddings, which condition the generator together with textual prompts. The semantic consistency discriminator calculates text-image cosine similarity based on CLIP cross-modal embedding, jointly scoring lexical semantic consistency and symbol reconstruction accuracy. The generator employs AdaIN for knowledge-conditional modulation, and multi-scale training and PWC-Net optical flow regularization collaboratively optimize local details and inter-frame stability. To balance global realism and local cultural details., we adopt a multi-scale training schedule and an attention controller that dynamically allocates textures and color palettes to culture-critical regions. Temporal stability is further improved with a lightweight consistency loss.Experiments on a curated cultural corpus and human-corrected pairs report cultural semantic fidelity of 89.7%–92.1%, relation compliance of 86.9%–88.6%, inter-frame SSIM under fast transitions of 0.782, and optical-flow smoothness of 0.751. A vocational Chinese-language classroom case study shows higher cultural expressiveness, learner engagement, and motion comfort compared with diffusion-based and CLIP-guided baselines, with ablations confirming the contributions of KG constraints, the semantic discriminator, and attention control. The framework is plug-and-play for XR/education/museums, and we provide KG schemas, code, and evaluation scripts to support reproducible deployment.

Povzetek: Članek predstavlja model umetne inteligence za ustvarjanje kulturno verodostojnih virtualnih prizorov, ki izboljšuje realizem, semantično skladnost in uporabniško izkušnjo v izobraževalnih ter XR okoljih.

1 Introduction

Chinese culture carries profound philosophical thoughts and aesthetic paradigms. Its inheritance in higher vocational Chinese language education urgently needs to break through the single path of text interpretation [1-2]. Integrating virtual scene construction into teaching can activate the historical context and emotional structure behind the language, enabling learners to understand the contextual logic through embodied perception [3]. Generative adversarial networks have the ability to generate cross-modal information from semantics to vision, providing a technical support for the dynamic reproduction of cultural imagery [4-5]. Through algorithm-driven cultural space reconstruction, not only does it expand the expressive dimension of Chinese language teaching, but it also builds a bridge between knowledge cognition and situational experience, promoting the evolution of traditional cultural education from "reading and knowing" to "understanding", which has profound teaching innovation significance.

Current generative model-based teaching scenarios suffer from multiple biases, primarily manifesting in semantic drift and symbol misplacement. While some systems can render realistic surface textures and lighting effects when generating images of ancient buildings or figures, key cultural elements such as clothing patterns, ritual vessel forms, and spatial orientation often exhibit chronological or regional mismatches [6-7]. For example, a Song Dynasty literati scene might feature furniture styles popular only during the Ming Dynasty, or ethnic minority totems might be arbitrarily embedded in the ritual spaces of the Central Plains. These biases are not accidental technical noise, but rather stem from a lack of structured knowledge constraints in the training data. Existing methods often rely on large-scale image-text pairs for end-to-end learning. While optimizing for pixel-level realism, these models struggle to capture deeper cultural logic [8-9]. Furthermore, interaction design often relies on click-to-browse or perspective switching, failing to dynamically adjust the content presentation level based on the learner's cognitive rhythm and focus [10-11]. Although users are immersed in a three-dimensional environment, they remain passively receptive and lack deep engagement with cultural details. A deeper problem lies in the fuzzy

mapping between generated results and course objectives, making it impossible to accurately serve the deepening of understanding of specific knowledge points [12-13]. Some platforms emphasize technical indicators of immersion, such as frame rate, resolution, and degrees of freedom, but ignore the integrity and logical coherence of cultural information transmission. The resulting experience appears rich, but in reality, the information is fragmented and cannot support systematic knowledge construction.

Early research on digital humanities teaching focused on the construction of multimedia resource libraries, using a combination of images and text to assist classroom teaching. Subsequently, virtual simulation technology was introduced, and scholars tried to manually build typical cultural scenes, such as academies, markets, and sacrificial ritual sites, using modeling software to achieve preliminary spatial reproduction [14-15]. This type of method ensures a high degree of cultural accuracy, but the development cycle is long, the cost is high, and it is difficult to adapt to diverse teaching needs. To improve flexibility, some teams turned to using generative models to automatically synthesize scene content. Some studies used conditional diffusion models to drive image generation with text descriptions, and made some progress in context restoration [16-17]. Other work combined style transfer technology to apply classical painting aesthetics to modern rendering processes to enhance visual beauty [18]. In recent years, generative adversarial networks have attracted attention for their powerful ability to generate details. Experiments have used them to visualize the artistic conception of ancient poetry and generate corresponding images by encoding the semantics of the poems [19-20]. Although these attempts have expanded the boundaries of technology, they still face core bottlenecks: the generation process lacks an external knowledge verification mechanism, and the model only splices visual elements based on statistical laws, which cannot ensure that the relationship between cultural entities conforms to historical facts. Worse still, some sacrifice factual evidence in pursuit of artistic expression, resulting in a "beautiful but unrealistic" outcome. Overall, existing research has yet to find an effective balance between authenticity and automation, with most approaches still treating cultural understanding as a pre-set input rather than a dynamic variable involved in generative decision-making.

To address the issue of generated content deviating from its cultural roots, some studies have begun exploring knowledge-enhanced generation architectures. One study embedded ontology models into variational autoencoders, leveraging concept hierarchies to guide image semantic layout, improving the rationality of pattern placement in restoration tasks[21-22]. Another approach employed graph neural networks to extract relevant features and inject them into the generation process as prior information, improving the logical consistency of artifact combinations[23-24]. In the interactive dimension, hierarchical response mechanisms[25-26] are used to regulate information density in virtual environments, dynamically loading annotation content based on user dwell time and gaze trajectory to avoid information overload. Furthermore, attention guidance strategies have demonstrated their effectiveness in accurately controlling key regions in medical image generation, suggesting the possibility of focusing on cultural elements[27-28]. Although these approaches demonstrate the potential for knowledge fusion and interaction optimization, they still face limitations in practical applications. Knowledge graphs are often implemented in the form of static rules, making them difficult to update and unable to cover complex contexts. Attention mechanisms are typically based on visual saliency calculations, ignoring the distribution of semantic importance weights. Hierarchical interaction designs often rely on predefined paths and lack real-time linkage with the generation system. These issues keep knowledge constraints superficial, failing to deeply influence the generative decision chain. This paper proposes a generative adversarial framework that integrates graph embedding constraints with semantic consistency judgment. This framework encodes cultural knowledge into computable vectors and continuously calibrates the output during the generator-discriminator game. Furthermore, it constructs an interactive feedback loop based on cognitive hierarchies, enabling virtual scenes to not only appear realistic but also be learned deeply.

This research aims to build a virtual Chinese language teaching system for higher vocational education that combines cultural fidelity with pedagogical adaptability, focusing on overcoming the dual challenges of semantic misalignment in generated content and a superficial immersive experience. The innovation lies in deeply coupling the structured representation of knowledge graphs into the training mechanism of a generative adversarial network, enabling the model to internalize cultural logic while learning visual distributions. Specifically, by mapping traditional cultural entities and their relationships into graph embedding vectors and jointly inputting them into the generator along with natural language descriptions, a semantically driven image synthesis pathway is implemented. A semantic consistency assessment module is added to the discriminator, leveraging a pre-trained language model to quantify the semantic distance between the generated image description and the original cultural text, generating a feedback signal of cultural authenticity that goes beyond visual realism. In terms of training strategy, a multi-scale optimization mechanism is employed to simultaneously enhance visual quality and symbolic accuracy at both the local texture and overall composition levels, ensuring that clothing patterns, architectural regulations, and spatial furnishings are consistent with the specific historical context. An attention mechanism is introduced to weight key cultural elements, dynamically adjusting color distribution and material details to improve the accuracy of reproducing core features. A layered response structure is designed at the interaction level, adaptively adjusting the granularity of information presentation based on learner behavioral data. This supports progressive exploration, from macro-level atmospheric perception to micro-level symbolic interpretation. The entire system not only enables the automated generation of high-quality

virtual scenes but, more importantly, establishes a semantic feedback loop via discriminator embedding between cultural accuracy and teaching effectiveness, providing a scalable and verifiable technical paradigm for the digital transformation of higher vocational Chinese language courses. The key contributions of this paper are as follows: (1) We propose the KG-GAN framework, which for the first time deeply integrates knowledge graph structural constraints into GAN's generation and discrimination mechanisms; (2) We design a semantic consistency discriminator and attention controller to achieve precise generation and dynamic allocation of cultural details; (3) We establish a multi-scale temporal evolution and hierarchical interaction mechanism to ensure logical coherence and teaching adaptability in dynamic scenarios; (4) We demonstrate the method's significant advantages in cultural fidelity, temporal stability, and immersion in vocational Chinese language teaching contexts.

1 Virtual cultural scene generation architecture based on semantic constraints

Figure 1 systematically presents the overall technical framework for constructing a virtual traditional cultural scene for vocational Chinese language education driven by a generative adversarial network. The knowledge graph module forms a structured cultural semantic network through multi-source text fusion and entity relationship extraction. The generator receives text and graph embedding vectors and produces culturally accurate scene images through multi-scale synthesis. The discriminator employs a dual-path architecture to simultaneously evaluate visual authenticity and semantic consistency. The interaction layer employs a three-level cognitive logic, leveraging natural language parsing to achieve a progressively immersive experience. These modules form a closed-loop optimization mechanism. The knowledge graph provides cultural constraints for generation and interaction, the discriminator provides feedback on semantic fidelity, and the interaction layer dynamically adjusts content presentation, effectively addressing the issues of cultural distortion and superficial experience.

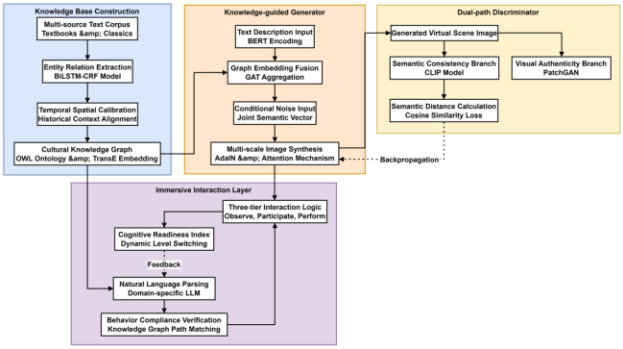

Figure 1: The technical framework for constructing virtual traditional cultural scenes driven by generative adversarial networks.

2.1 Knowledge graph-driven cultural semantic modeling

2.1.1 Traditional cultural entity relationship extraction and structural modeling based on multi-source text fusion

In order to achieve a systematic expression of traditional cultural elements in the Chinese language curriculum for higher vocational education, we first collected original corpus from current unified textbooks, annotations to classic ancient books, and cultural research literature supported by the National Social Science Fund to construct a domain-specific corpus. A hybrid extraction strategy combining rule-driven and deep learning is adopted to perform dual analysis of explicit statements and implicit associations in the text. For core entity categories such as characters, allusions, ritual systems, and artifact shapes, a labeling framework based on the joint modeling of dependency syntax analysis and named entity recognition is designed. The BiLSTM-CRF model is used to complete preliminary entity boundary identification, and an external dictionary is introduced to enhance the robustness of recognition of ancient proper nouns. On this basis, by defining domain-specific trigger word templates and semantic role labeling rules, complex relationships between entities such as spatiotemporal co-occurrence, behavioral giving and receiving, and symbolic meaning are extracted. For long-distance dependencies across sentences, a relational classification network based on the attention mechanism is used to calculate the semantic association strength of candidate entity pairs within the context window:

$$s_{ij}=\mathbf{v}^{T}\tanh\left(\mathbf{W}_{\mathrm{h}}\left[\mathbf{h}_{i};\mathbf{h}_{j}\right]+\mathbf{b}_{\mathrm{h}}\right) \quad (1)$$

Where represents $s_{ij}$the confidence score for the relationship between $\mathbf{h}_{i}$,$\mathbf{h}_{j}$entities $j$and , $i$is the hidden state vector of the corresponding entity in the context encoding layer, $\mathbf{W}_{\mathrm{h}}$and $\mathbf{b}_{\mathrm{h}}$are the trainable parameter matrix and bias term, and v is the attention weight vector. This score is used to select high-confidence triplets for inclusion in the initial screening set of the knowledge graph. Subsequently, all time nodes are normalized and calibrated based on the historical chronological coordinate system established in classics such as the "Book of Rites" and the "Book of Zhou Rites," establishing a unified timeline reference system. This is combined with geographical data to complete the semantic mapping of spatial coordinates, ensuring the logical closure of the "person-event-place" triad in both temporal and spatial dimensions.

2.1.2 Graph Ontology Construction and Semantic Consistency Verification Mechanism

After extracting the basic triples, a hierarchical ontology model is constructed to clarify inheritance, parallelism, and constraint relationships between categories. Using OWL (Web Ontology Language) as a formal description language, high-level concept nodes such as "Confucian Rituals," "Poetic Imagery," and "Historical Allusions" are defined. These concepts are then refined down to the specific instance level through subclassing, forming a tree-like topology from abstract categories to concrete entities. Edge relationship types are

strictly limited to a set of predefined semantic predicates, including "occurs in" (temporal association), "belongs to" (category affiliation), "symbolizes" (value expression), and "use" (artifact function), to avoid semantic ambiguity or redundant connections. To improve the internal logical consistency of the graph, a reasoning engine based on description logic is introduced to perform consistency checks and infer implicit relationships. The following constraint is set: $\forall x \in$Person, if hasRank $(x,r) \wedge r <$minRankForCeremony (c), then $\neg$participatesIn (x,c). This axiom states that if a person's rank is below the minimum qualification for participation in a specific ritual, they cannot participate in the ritual activities, thereby preventing the generation of scene configurations that do not conform to ritual norms. In addition, the TransE embedding method is used to map entities and relationships in the graph into a continuous vector space, and the degree of semantic proximity is measured by geometric distance. The objective function is optimized through negative sampling:

$$L=\sum_{(h,r,t)\in T}[\gamma+\|\mathbf{e}_h+\mathbf{r}_r-\mathbf{e}_t\|_2-\|\mathbf{e}_{h'}+\mathbf{r}_r-\mathbf{e}_{t'}\|_2]_+ \quad (2)$$

Where $T$is the set of positive triples, $(h',t')$is the negative replacement pair, $\gamma$is the margin threshold, $\mathbf{e}_h$,$\mathbf{e}_t$is the head and tail entity embeddings, $\mathbf{r}_r$and is the relationship direction. This process not only strengthens the semantic coherence of the graph itself but also provides a differentiable knowledge representation input channel for subsequent generative models. The resulting structured knowledge base features a dynamic update interface, allowing teachers to inject new knowledge points based on teaching progress, ensuring that the graph content iterates synchronously with curriculum evolution.

2.2 Semantic embedding enhanced generative adversarial network training

2.2.1 Generator conditional input construction and semantic guidance mechanism based on graph embedding fusion

To achieve accurate mapping of cultural elements in the visual generation process, a generator input structure centered on multimodal conditional vectors is constructed. Entities and their relationships extracted from the knowledge graph are encoded into fixed-dimensional graph embedding vectors using the TransE or RotatE methods, ensuring that semantically similar cultural concepts maintain geometric proximity in the vector space. Specifically, for key cultural nodes involved in the input text, the set of directly adjacent triples in the knowledge graph is retrieved, and a graph attention mechanism is used to aggregate local structural information to generate context-aware entity representations:

$$\mathbf{g}_e = \sum_{k\in N(e)} \alpha_{ek}\cdot\sigma\left(\mathbf{W}_r(\mathbf{h}_k+\mathbf{r}_{rk})\right) \quad (3)$$

Where $g_e$represents $e$the aggregate embedding of the target entity, $N(e)$its neighborhood triplet index, $\mathbf{h}_k$is the initial embedding of the neighboring entities, $\mathbf{r}_{rk}$is the correspondence vector, $\mathbf{W}_r$is the learnable transformation matrix, $\sigma$and is the nonlinear activation function. Attention weights α_{ek}, calculated using dot-product attention, measure the contribution of different neighboring information to the current generation task. The resulting graph embedding vector is normalized and then channel-wise concatenated with a natural language description encoded using the BERT (Bidirectional Encoder Representations from Transformers) model to form a joint conditional input tensor. This tensor serves as the initial noise modulation signal for the deep convolutional generative network. Semantic priors are injected layer-by-layer through Adaptive Layer Normalization (AdalN) between deconvolutional layers, enabling texture generation, object layout, and color configuration to be governed by structured cultural knowledge. In particular, in scenarios involving rituals or hierarchical differences in character, the identity attributes and spatial sequence constraints carried by the graph embeddings are explicitly decoded into the spatial attention map of the feature map, ensuring that key visual elements such as clothing style and standing order conform to historical norms.

2.2.2 Discriminator dual-path evaluation architecture and semantic consistency feedback mechanism

In order to break through the limitation of traditional discriminators that only focus on pixel-level authenticity, a dual-branch discriminant structure is designed that includes visual authenticity judgment and semantic consistency verification. The main path follows the PatchGAN framework to judge the authenticity of local texture details of the generated image and maintain the visual authenticity of high-resolution output. The newly added semantic consistency branch is responsible for evaluating the deep semantic alignment between the generated content and the original cultural description. In the specific implementation, the CLIP (Contrastive Language-Image Pre-Training) pre-trained multimodal model is first used to extract the semantic encoding of the generated image $\mathbf{v}_i$and the semantic encoding of the original text description respectively t_d. The two belong to the shared cross-modal embedding space. On this basis, the cosine similarity measurement mechanism is introduced to optimize the matching level of the two at the cultural connotation level:

$$S_{sem}=1-\frac{\mathbf{v}_i\cdot\mathbf{t}_d}{\|\mathbf{v}_i\|\|\mathbf{t}_d\|} \quad (4)$$

This distance value $S_{sem}$is used as a semantic deviation indicator to participate in the construction of the discriminant loss function. The total loss of the discriminator is composed of two weighted parts:

$$L_D=L_{adv}+\lambda L_{sem} \quad (5)$$

Where $L_{adv}$is the standard adversarial loss, $\nu L_{sem}=S_{sem}$is the semantic deviation penalty term, and the hyperparameter λcontrols the optimization intensity of knowledge fidelity. This mechanism enables the discriminator to not only reject obviously distorted images, but also to identify content that is visually reasonable but culturally misplaced—such as inappropriate combinations of utensils in specific rituals,

or implicit errors such as unethical postures of characters. During backpropagation, the semantic loss gradient is passed back to the generator via the shared feature layer, driving it to adjust its latent space sampling strategy and pending activation pattern, thereby gradually converging on a generation trajectory that is both visually credible and culturally compliant during iterative training. The entire training process is optimized end-to-end on a dataset of typical Chinese language passages for higher vocational education, ensuring that the generated results are tightly coupled with the curriculum knowledge system, achieving closed-loop calibration from semantic understanding to visual reproduction. To ensure GAN's training stability across diverse text and graph input types, this framework employs spectral normalization to constrain discriminator weights while incorporating gradient penalty terms to prevent pattern collapse. The integration of graph embedding and text encoding normalization effectively mitigates conditional input distribution shifts, thereby enhancing the generator's convergence consistency.

Table 1: Core parameter configuration

Parameter Name	Value	Description
Generator Learning Rate	0.0002	Step size for generator optimization
Discriminator Learning Rate	0.0002	Step size for discriminator optimization
Batch Size	16	Number of samples per training iteration
Graph Embedding Dimension	256	Dimension of knowledge graph embeddings
Text Encoding Dimension	768	Feature dimension from BERT text encoder
Combined Input Dimension	1024	Concatenated dimension of dual-modal input
Number of AdaIN Layers	6	Adaptive normalization layers in generator
PatchGAN Patch Size	32×32	Local discrimination resolution
CLIP Embedding Dimension	512	Cross-modal alignment space dimension
Semantic Loss Weight	0.8	Weight for semantic consistency term
Adam β_1	0.5	Decay rate for first-moment estimation
Adam β_2	0.999	Decay rate for second-moment estimation
Graph Attention Heads	4	Number of parallel attention heads

Table 1 lists the core parameter configurations for the semantic embedding-enhanced generative adversarial network training process, covering key hyperparameters related to model optimization, structural design, and multimodal fusion. These include settings for the learning rate and momentum of the generator and discriminator, batch processing size, the dimensional composition of the conditional input vector, the number of normalized modulation layers, control of discriminant granularity, cross-modal alignment space settings, and attention mechanism structural parameters. All parameters were specifically selected based on the specific characteristics of the Chinese language and cultural scene generation task in vocational colleges. This ensures that semantic information from the knowledge graph and text descriptions are effectively integrated into the image generation process, achieving a coordinated optimization between visual realism and cultural consistency, supporting the output of high-quality content for subsequent immersive teaching applications.

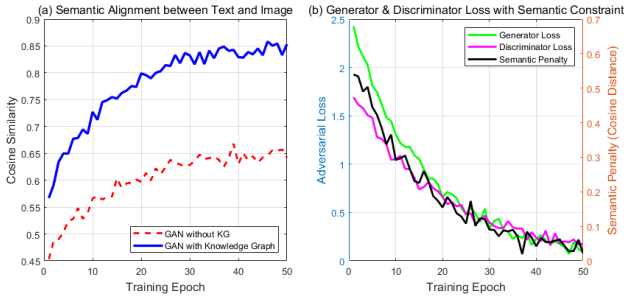

Figure 2 Cosine similarity and loss convergence trend

Figure 2(a) shows the cosine similarity between the text description and the generated image over 50 training epochs (Training Epoch on the horizontal axis, Cosine Similarity on the vertical axis). The red dashed line represents the GAN without the knowledge graph, and the blue solid line represents the GAN with the knowledge graph constraint. As can be seen, the semantic match of both models steadily improves with each training epoch, but the model with the knowledge graph ultimately reaches around 0.85, while the model without the knowledge graph only reaches around 0.65. The curves rise smoothly with almost no sharp fluctuations, consistent with the gradual optimization of semantic embeddings during actual training. This indicates that the knowledge graph constraint effectively enhances the generator's grasp of cultural semantics. Figure 2(b) uses two vertical axes: the left axis represents the adversarial loss between the generator and discriminator (green and purple solid lines), and the right axis represents the semantic penalty (black solid line, which measures the semantic deviation between the generated image and the text description). As can be seen, the generator and discriminator losses steadily decrease during training, eventually converging to lower values, and the semantic penalty also steadily decreases to near zero. This shows that the semantic consistency branch of the discriminator can continuously provide feedback to the generator, prompting it to optimize the generated content in terms of both visual authenticity and cultural semantics, thereby achieving a "true and accurate" generation effect.

2.3 Multi-scale cultural scene synthesis and dynamic evolution mechanism

2.3.1 Knowledge-driven sequential scenario generation sequence modeling and structured evolutionary control

In order to achieve the dynamic restoration of traditional cultural activities in virtual space, static image generation is extended to the time dimension, and a multi-frame generation architecture with semantic coherence as the core is constructed. First, based on the plot development context described in the teaching text, key event nodes are extracted and mapped to the action links in the knowledge graph to form a scene evolution path with causal dependencies and temporal constraints. The generation conditions of each frame of the image not only include the text description of the current moment, but also introduce the cultural state vector of the previous frame as context memory to ensure that the identity of the characters, spatial configuration and ritual process maintain logical continuity. Specifically, the scene state latent variable $\mathbf{s}_t$is defined to represent the cultural stage of the t-th frame. It is extracted from the output of the previous frame through a lightweight encoder and gated fused with the current instruction text and the embedding vector of the corresponding node in the knowledge graph:

$$\mathbf{c}_t=\mathbf{z}_t\odot\sigma\left(\mathbf{W}_g\left[\mathbf{s}_{t\text{-}1};\mathbf{e}_{\text{event}_t}\right]\right)+(1\text{-}\mathbf{z}_t)\odot\mathbf{t}_t \quad (6)$$

Where $\mathbf{c}_t$is the joint conditional vector of the final input generator, $\mathbf{z}_t$is the learnable gating weight, $\mathbf{s}_{t\text{-}1}$represents the state encoding of the previous moment, $\mathbf{e}_{\text{event}_t}$is the semantic embedding of the current event in the knowledge graph, $t_t$is the original text encoding, $\mathbf{W}_g$is the transformation matrix, and $\sigma$is the Sigmoid function. This mechanism effectively prevents character dislocation or process jumps caused by local description ambiguity, ensuring the stability of the overall narrative structure. Furthermore, a consistency check module is implemented at key turning points in the generated sequence. This module utilizes a pre-trained language model to determine the semantic implications of the description text corresponding to adjacent frames. If a logical break is detected, a resampling mechanism is triggered to readjust the generation parameters of the intermediate frames, ensuring the integrity of the cultural behavior chain.

2.3.2 Multi-scale Visual Continuity Preservation and Cross-frame Style Co-optimization

While ensuring correct semantic evolution, a method combining optical flow guidance and latent space interpolation is employed to enhance motion smoothness and visual coherence between frames. For pixel-level transitions between adjacent generated frames, the PWC-Net architecture is deployed to estimate forward and backward optical flow fields, capturing fine-grained motion trajectories of character pose changes, object displacement, and ambient light and shadow evolution. The resulting optical flow map serves as a regularization term to constrain the generator's feature updates, ensuring that the spatial distribution of the same entity at different moments conforms to the laws of physical motion. Furthermore, a latent vector interpolation strategy is introduced to linearly or spherically interpolate the noise input of consecutive frames within the latent space to avoid abrupt content jumps. Furthermore, to maintain overall aesthetic consistency, a cross-frame style synchronization mechanism based on AdaIN is designed: style statistics (channel mean and variance) are extracted from the generated results of the first frame and transferred to the intermediate feature representations of subsequent frames through a differentiable normalization layer, achieving unified control of color tone, brushstroke texture, and compositional rhythm. To quantify the quality of visual continuity, a perceptual difference metric between frames is defined:

$$D_{\text{temp}}=\frac{1}{T\text{-}1}\sum_{t=1}^{T\text{-}1}\|\phi(I_t)\text{-}\phi(I_{t+1})\|_2 \quad (7)$$

Where D_{temp}represents the temporal smoothness of the entire sequence, $I_t$is $t$the frame image, and is the output symbol of a deep activation layer $\phi(\cdot)$in the VGG (Visual Geometry Group) network, which is used to capture the degree of change in high-level semantic structure. This metric is incorporated into the generator's optimization objective, forming a multi-objective function along with the adversarial loss and semantic consistency loss, driving the model to strike a balance between visual authenticity and dynamic fluidity in cultural logic. The entire synthesis process supports on-demand adjustment of frame rate and transition speed to accommodate immersive playback requirements at varying teaching tempos, thereby achieving the generation of cultural scene sequences that are both faithful to the historical context and possess a natural, dynamic aesthetic.

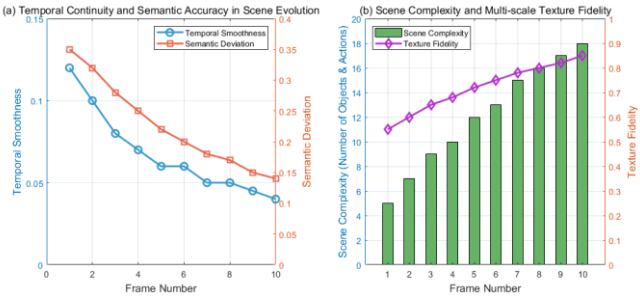

Figure 3: Multi-scale dynamic evolution and changes in visual quality indicators

Figure 3 illustrates the multi-scale dynamic evolution and changes in visual quality indicators during the generation of a virtual cultural scene for a vocational Chinese language course. The left sub-figure shows the time frame number as the horizontal axis, the left vertical axis represents the inter-frame visual continuity indicator, and the right vertical axis represents the semantic deviation of key events. The data shows that as the frame sequence progresses, Temporal Smoothness decreases from 0.12 to 0.04, indicating a significant improvement in inter-frame motion transition smoothness. Simultaneously, Semantic Deviation decreases from 0.35 to 0.14, reflecting a gradual increase in semantic consistency between the generated content and the original text, validating the effectiveness of the knowledge

graph constraint and semantic consistency judgment mechanism. The right sub-figure also shows the frame number as the horizontal axis, the scene complexity (including the number of activities, costumes, and artifact types) as the left vertical axis, and the multi-scale texture fidelity as the right vertical axis. The data show an upward trend, with complexity increasing from 5 to 18 and texture fidelity improving from 0.55 to 0.85, demonstrating that the generator effectively enriches cultural elements and optimizes local and global visual details during the multi-frame evolution process. The overall diagram intuitively demonstrates that the system achieves simultaneous optimization of scene dynamic continuity and visual fidelity while maintaining cultural semantic accuracy, providing quantitative support for multi-scale cultural scene synthesis.

2.4 Design and behavior mapping of layered interaction interfaces

2.4.1 Structured modeling and cognitive load adaptation mechanism of the three-level interaction hierarchy

In order to achieve a coordinated improvement in user cognitive depth and operational freedom during the immersive learning process, a hierarchical interactive architecture based on cognitive progressive logic is constructed. The system sets three levels: observation, participation, and interpretation, which correspond to the three learning states of passive perception, active intervention, and identity substitution, and realizes gradual switching between levels through permission control and interface guidance. At the observation level, users browse the generated cultural scenes from a third-person perspective. The interactive interface only provides hotspot marking and lightweight annotation pop-up functions to avoid information overload interfering with situational immersion. After entering the participation level, the system activates the semantic response area of the operable object, allowing users to select specific objects, people, or spatial nodes through gestures or pointers to trigger the action simulation module associated with them. At the interpretation level, users need to complete the role binding process, and their identity information is mapped to the specific behavior subject in the knowledge graph. All subsequent behaviors are subject to the social hierarchy, ritual authority, and language style of the role. Dynamic transition thresholds are set between each level, and the current cognitive readiness index is calculated based on the user's stay time, interaction frequency, and question answering accuracy η_t:

$$\eta_t=\omega_1\cdot\frac{1}{T}\sum_{\tau=t-T}^{t}a_\tau+\omega_2\cdot p_{\text{acc}}+\omega_3\cdot l_{\text{dur}} \quad (8)$$

Where $a_\tau$represents $T$the number of valid interactions in the past time steps, $p_{\text{acc}}$is the accuracy l_{dur}rate of the most recent round of knowledge quizzes, is the normalized average dwell time at the current level, $\omega_1,\omega_2,\omega_3$and is the weight coefficient set based on the teaching objectives. When $\eta_t$the preset threshold is exceeded, the system automatically prompts users to upgrade their interaction permissions, ensuring they have sufficient background understanding before entering a more advanced learning mode. This mechanism effectively prevents the frustration of shallow explorers jumping directly into complex tasks while encouraging deeper learners to gain a richer cultural experience.

2.4.2 Natural Language-Driven Behavioral Semantic Parsing and Compliance Mapping Strategy

In order to achieve precise docking between user intentions and virtual environment responses, an end-to-end semantic parsing-behavior matching pipeline is designed. The natural language instructions input by the user are first subjected to a syntactic-semantic joint analysis by a domain-fine-tuned Chinese pre-trained model (such as ChatGLM-6B), and the action verbs, objects of action, and modifying and limiting components are extracted and converted into standardized predicate forms. Subsequently, the system searches the knowledge graph for behavior nodes that match the predicate structure, locates the etiquette category, applicable field, and execution subject qualifications to which it belongs. If the user identity meets the preconditions, the animation synthesis engine is started to call the preset action sequence; otherwise, alternative suggestions that comply with etiquette norms are returned. To ensure the accuracy of semantic parsing, a contextual disambiguation mechanism based on the knowledge graph path is introduced to calculate the semantic affinity between the candidate behavior node and the current scene context:

$$\kappa_b=\exp\left(-\beta\cdot d_{\text{path}}(b,s_{\text{curr}})\right) \quad (9)$$

Where $\kappa_b$represents $b$the activation confidence of the behavior node d_{path}in the current scenario , s_{curr}represents the shortest path length between it and the main event node of the scenario in the knowledge graph, $\beta$and represents the attenuation coefficient, which is used to suppress cross-context mismatches. Ultimately, $\kappa_b$the highest legal behavior is selected as the execution target. For tasks involving dialogue interaction, the system calls a response generation module based on a combination of template filling and semantic rewriting to output text feedback that is consistent with the character's identity and the language of the times, and simultaneously drives the update of the virtual character's lip shape, expression, and posture parameters. While ensuring operational flexibility, the entire interactive closed loop strictly adheres to the behavioral paradigm of traditional culture, enabling learners to internalize the rules of etiquette and value logic in embodied practice, achieving a cognitive leap from "seeing" to "practicing."

2 Quantitative comparison of virtual teaching scene generation quality and experience effectiveness

3.1 Experimental data

The experimental dataset is constructed from standard texts from vocational Chinese language courses and accompanying classic texts, covering three thematic

categories: festival rituals, ritual and music systems, and ancient architectural forms. By systematically combing through annotations and authoritative textbooks from classic texts such as the Book of Songs, the Book of Rites, and the Records of the Grand Historian, key cultural elements were extracted, resulting in 12,843 structured triples of entities, including characters, objects, actions, and spatial configurations, forming the foundation of the ontological knowledge graph. The visual generation training set consists of 6,720 high-precision line drawings and historically restored images drawn by professional artists, each annotated with fine-grained semantic labels and corresponding text descriptions. The dynamic sequence test set selects typical cultural processes such as "sacrificial rites," "academic lectures," and "flowing wine cups on a winding stream," constructing standard animation samples lasting 30 seconds and at a frame rate of 25 fps. The interactive experiment recruited 30 vocational college students, who completed standardized tasks while simultaneously collecting physiological signals and subjective feedback. All text-image pairs were cross-validated by three cultural and historical experts to ensure semantic accuracy and contemporary compliance, providing reliable data support for model training and multi-dimensional evaluation. The experiment was conducted on a server equipped with 4 NVIDIA A100 GPUs (80GB each), with the training process taking approximately 68 hours.

3.2 Analysis of the effect of knowledge graph semantic modeling

For the relationship extraction task, an annotated corpus containing multiple types of semantic relationships is constructed. Relationships in the corpus are identified and classified by combining a pre-trained language model with a knowledge graph embedding method. Statistics are collected for the extraction results of each type of relationship, and the average confidence and standard deviation are calculated to reflect the stability and reliability of the model for different relationship types. For semantic modeling between entities, embedding algorithms such as TransE are used to map cultural entities into a low-dimensional vector space. The semantic proximity between entities is calculated based on Euclidean distance to form a distance matrix. The entire evaluation process uses cross-validation and error analysis to ensure a reasonable data distribution and avoid distorted results due to biased samples.

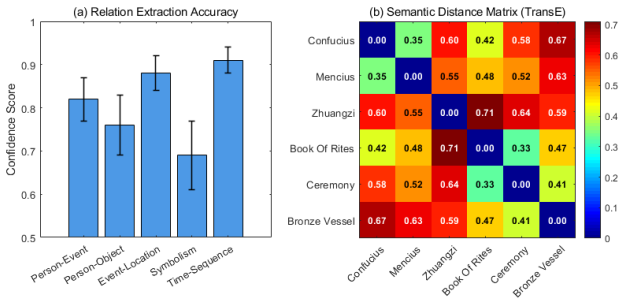

Figure 4: Knowledge graph semantic modeling results

Figure 4 shows the results of knowledge graph-driven cultural semantic modeling. Figure 4(a) shows the relationship extraction accuracy statistics, with the horizontal axis representing the five relationship types and the vertical axis representing the model's average confidence score for the relationships. The data shows that Time-Sequence (0.91) and Event-Location (0.88) have the highest confidence scores, indicating that the model is most reliable in processing temporal logic and spatial mapping. Person-Event (0.82) and Person-Object (0.76) are at intermediate levels, while Symbolism is only 0.69, indicating that abstract symbolic relationships still face difficulties in recognition, revealing the impact of semantic hierarchy complexity on model performance. Figure 4(b) shows the TransE semantic distance matrix, which displays the Euclidean distances of six cultural entities in the embedding space. The distances between Confucius and Mencius (0.35) and between the Book of Rites and Ceremonies (0.33) are relatively small, reflecting the high correlation between their ideas and institutions. The distances between Confucius and Zhuangzi (0.60) and between Zhuangzi and bronze vessels (0.59) are relatively large, reflecting the differences between philosophical thought and material objects. Overall, the data reveals that the model has high modeling accuracy in specific entities and spatiotemporal relationships, but still needs to be optimized in abstract symbols and cross-category relationships, verifying the significant impact of semantic level differences in knowledge graph modeling on the results.

Based on the annotated test corpus, entity relationship extraction was performed using the proposed hybrid model (a BiLSTM-CRF network combined with a rule-based template and attention mechanism for relation classification) and a baseline model (a pure BiLSTM-CRF model). Model performance was recorded on a per-epoch basis by calculating precision and recall, and synthesizing them to form the F1-score. High-confidence triples predicted by the model in each round were manually verified and added to the graph, and their cumulative number was counted. Performance metrics and graph size data were collected and analyzed simultaneously during the same training rounds. The proposed model refers to a hybrid extraction architecture that combines rule-based priors with an attention mechanism, while the baseline model uses a standard approach using only neural network end-to-end learning.

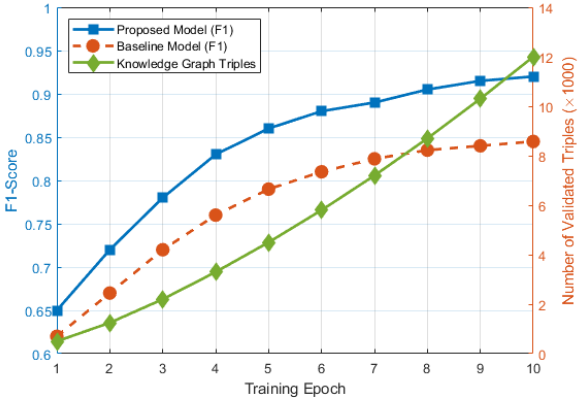

Figure 5: Entity relationship extraction performance and knowledge graph growth

Figure 5 uses two vertical axes to clearly compare the performance of the entity relationship extraction model with the growth of the knowledge graph itself. The left vertical axis represents the F1-Score of the model's extraction performance, ranging from 0.6 to 1.0; the right vertical axis represents the number of triples incorporated into the knowledge graph after verification (in thousands), ranging from 0 to 14,000; and the horizontal axis represents the epochs of model training. The left axis contains two curves: the "Proposed Model (F1)" marked by the blue solid square represents the performance of the hybrid extraction strategy proposed in this paper. Its F1-Score steadily improves from an initial 0.65 with increasing training epochs, ultimately converging to a high level of 0.92, demonstrating that the model effectively learns complex semantic patterns in the text; the "Baseline Model (F1)" marked by the orange dotted circle represents the performance of the baseline model. Its performance increases relatively slowly and its final performance is significantly lower than that of our model, highlighting the advantages of our approach in solving the problem of ancient proper nouns and long-distance dependencies. The data on the right axis is the "Knowledge Graph Triples" curve, marked by green solid diamonds. This curve shows that as model performance improved and the verification process progressed, the number of valid triples in the knowledge graph rapidly increased, ultimately reaching 12,000. This data demonstrates that high-performance and stable extraction models are fundamental to building large-scale, high-quality knowledge graphs, and their growth trends are positively correlated. The method proposed in this article can efficiently and reliably support subsequent cultural semantic modeling.

3.3 Quantitative comparison of cultural semantic fidelity

The evaluation metrics used are Cultural Entity Accuracy (CEA), which measures the percentage of correctly identified traditional cultural elements in generated scenarios, and Relationship Compliance Index (RCI), which measures whether relationships between entities conform to the knowledge graph definition. Given the same text prompt, Cultural Entity Accuracy (CEA): In the generation scenario, three domain experts independently annotate all identifiable cultural entities (e.g., clothing, artifacts, architectural components). CEA is defined as the percentage of annotated entities that fully match the corresponding entity categories and attributes in the knowledge graph. Relation Compliance Index (RCI): For any two associated entities in the generation scenario, if their co-occurrence relationships (e.g., "use," "symbolize," "locate") align with the defined relationship types and directions in the knowledge graph, it is considered compliant. RCI is the percentage of compliant relationship pairs out of all detected relationship pairs. this paper tested three popular models: KG -GAN (Knowledge Graph-Generative Adversarial Network) , StyleGAN3 (Alias-Free Generative Adversarial Networks) , Diffusion Model, and CycleGAN, on topics related to festivals, rituals, and architecture.

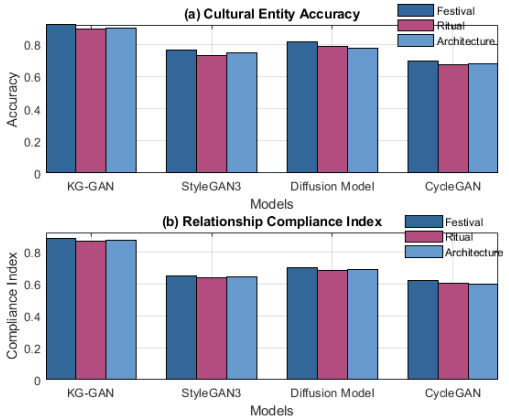

Figure 6: Cultural entity accuracy and relationship compliance index

Figure 6 compares the performance of different generative models using the dual-dimensional metrics of cultural semantic fidelity. The upper and lower subfigures show the distribution of cultural entity accuracy and relationship compliance index for three traditional cultural themes: festivals, rituals, and architecture. The overall trend shows that the proposed KG-GAN model significantly outperforms StyleGAN3, Diffusion Model, and CycleGAN in all categories, particularly in high-level semantic representation involving ritual norms and spatial logic. Its cultural entity accuracy ranges from 89.7% to 92.1%, and its relationship compliance index ranges from 86.9% to 88.6% . This is primarily due to the KG-GAN model deeply embedding the structural constraints of the knowledge graph into the generation process, ensuring that entity generation not only meets visual plausibility but is also dynamically calibrated to historical context and cultural norms. In contrast, other models rely on data-driven statistical models, which can easily lead to anachronisms and misplaced symbols. Compliance is particularly evident in artifact combinations and interactions between people, with noticeable declines in compliance. Furthermore, architectural scenes, due to their clear hierarchical forms and spatial order, showed particularly significant differences across models, further highlighting the necessity of knowledge-based guidance mechanisms in modeling complex cultural systems. This result validates the crucial role of deep integration of semantic priors and generative architecture in enhancing the credibility of virtual instructional content.

3.4 Visual coherence and temporal stability measurement

Structural Similarity Index Measure (SSIM) and Flow-based Motion Smoothness (FMS) are used to quantify the visual continuity and temporal stability of multi-frame scenes. The results generated by running the proposed KG-GAN alongside three popular models , StyleGAN3, Diffusion Model, and CycleGAN, on a 30-

second tutorial animation sequence were analyzed. The fluctuation ranges of these two metrics were measured under different tempos (slow narrative vs. fast transitions).

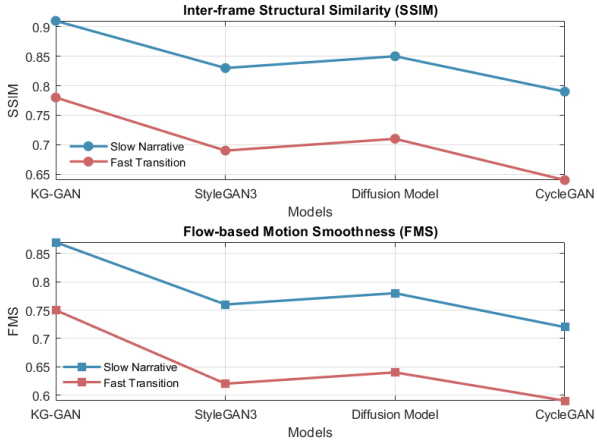

Figure 7: Visual coherence and temporal stability

Figure 7 compares the visual coherence and temporal stability of four generative models under different narrative rhythms. The upper and lower subfigures show the performance distribution of the inter-frame structural similarity index (SSIM) and optical flow motion smoothness (FMS), respectively. The overall trend shows that the proposed KG-GAN exhibits the best dynamic consistency under both rhythm conditions. Under the fast transition condition, the inter-frame structural similarity index is 0.782 and the optical flow motion smoothness is 0.751 . The fundamental reason for this is that KG-GAN, through knowledge graph-guided temporal state modeling, effectively constrains the magnitude of semantic transitions between adjacent frames, avoiding global structural fluctuations caused by local texture optimization. In contrast, models such as StyleGAN3 and the Diffusion Model, while performing well in single-frame quality, lack deep modeling of cultural behavioral logic and spatial evolution in multi-frame sequences, resulting in non-physical distortions such as sudden changes in character posture and jumps in object position. Furthermore, under the slow narrative condition, the differences between the models were relatively small, indicating that temporal redundancy can partially mask generation flaws. However, under the fast-switching condition, the limitations of temporal modeling were significantly amplified, highlighting the architectural superiority of KG-GAN in preserving dynamic semantics. This result demonstrates that generation mechanisms relying solely on visual statistical features are unable to meet the stringent requirements of spatiotemporal continuity for educational-level virtual scenes, and that cross-frame structured prior support is essential.

3.5 Cultural adaptability test of interactive response

Define Behavioral Match The former refers to the proportion of system responses after user actions that align with cultural norms, and the latter records the number of interactions that lead to logical contradictions or anachronisms. The performance of this paper's KG -GAN is compared with StyleGAN3, Diffusion Model, and CycleGAN under different pacing conditions (slow narrative/fast transition).

Table 2: Behavior matching rate and situational conflict rate

Model	Condition	Behavioral Match Rate (%)	Situational Conflict Rate (/min)
KG-GAN	Slow Narrative	94.7 ± 1.8	0.21 ± 0.06
	Fast Transition	92.3 ± 2.4	0.38 ± 0.09
StyleGAN3	Slow Narrative	78.5 ± 3.6	1.52 ± 0.21
	Fast Transition	73.1 ± 4.2	2.04 ± 0.33
Diffusion Model	Slow Narrative	75.9 ± 3.9	1.68 ± 0.25
	Fast Transition	70.4 ± 4.5	2.21 ± 0.37
CycleGAN	Slow Narrative	69.2 ± 4.7	2.43 ± 0.41
	Fast Transition	65.8 ± 5.1	2.87 ± 0.46

Table 2 presents the results of the cultural adaptability tests of different generative models in the interactive response phase, focusing on their ability to provide compliant feedback on user behavior and maintain contextual consistency under dynamic instructional pacing. The proposed KG-GAN maintains high stability under rapid transition conditions , with a behavior matching rate of 92.3% ± 2.4% and a contextual conflict rate of (0.38 ± 0.09) /min . This demonstrates that its knowledge-graph-driven behavior mapping mechanism effectively implements real-time verification of identity constraints, ritual rules, and spatiotemporal logic. In contrast, StyleGAN3, the Diffusion Model, and CycleGAN lack structured semantic support, and their system responses rely heavily on statistical correlations in the data distribution. This makes it difficult to identify implicit cultural taboos and hierarchical norms, leading to a high frequency of contextual conflicts. In fast-paced operations, user intent triggers intensive action sequences, and non-knowledge-guided models are prone to anachronisms such as role overreach and artifact misuse, exposing the fundamental limitations of the purely visual generative paradigm in understanding interactive semantics. Furthermore, while slow narrative can alleviate

the accumulation of conflict to a certain extent, it cannot compensate for the systematic biases caused by the lack of underlying logic. This result verifies that cultural intelligence interaction not only depends on response speed and visual reality, but also requires deep knowledge reasoning ability to ensure the historical legitimacy and educational seriousness of teaching behavior.

3.6 Subjective and objective dual measurement of immersion perception intensity

To comprehensively assess the impact of virtual traditional cultural scenarios on students' cognitive engagement and emotional involvement, a multimodal assessment system based on physiological signals and subjective reports was constructed. Within a standardized teaching experimental environment, participants wore portable physiological recorders to measure the low-frequency to high-frequency power ratio (LF/HF) of heart rate variability (HRV) , reflecting the autonomic nervous system's regulatory response to situational stimuli. Simultaneously, the degree of suppression of the occipital EEG alpha band (8–13 Hz) was measured as a neural indicator of attentional focus and psychological engagement. The subjective dimension was assessed using the NASA-TLX scale, which includes six sub-indicators, including mental demand and effort. The ITC-SOPI questionnaire measured three core immersion components: spatial presence, engagement, and sense of control. All tests were completed within the same class duration. The sample size (N=30) was determined by G*Power 3.1's prior power analysis (α=0.05, power=0.80, medium effect size f=0.25), meeting the minimum requirements for repeated measures ANOVA. All between-group comparisons were validated through Shapiro-Wilk normality test and Levene's homogeneity of variance test, followed by one-way ANOVA and Tukey post-hoc analysis. Both the α-suppression rate and ITC-SOPI score showed statistically significant improvements in the KG-GAN group compared to baseline ($p<0.01$).

Table 3: Subjective and objective dual-measurement of immersion perception intensity

Model	HRV (LF/HF)	α-Power Suppression (%)	NASA-TLX Score	ITC-SOPI Score
KG-GAN	1.98 ± 0.28	44.0 ± 5.8	9 0.5 ± 7.3	5.55 ± 0.56
StyleGAN3	1.97 ± 0.34	35.9 ± 6.6	8 5.5 ± 7.8	4.84 ± 0.63
Diffusion Model	1.97 ± 0.37	35.5 ± 7.0	8 6.3 ± 8.2	4.76 ± 0.66
CycleGAN	1.96 ± 0.41	32.0 ± 7.5	7 9.5 ± 8.6	4.44 ± 0.74

[Note: Data are composite means ± standard deviations; HRV units are dimensionless ratios; alpha wave suppression is based on baseline resting state; the maximum score for the NASA-TLX is 100, and the maximum score for the ITC-SOPI is 7. All physiological and subjective indicators were collected using a standardized experimental process. The results reflect the systematic impact of different generative models on learners' perception of immersion.]

Table 3 presents a comprehensive evaluation of the immersiveness of four generative models across multiple dimensions, encompassing both physiological responses and subjective experience. KG-GAN significantly outperforms the comparison models on key immersion metrics, including alpha wave suppression (44.0% ± 5.8%) and the ITC-SOPI score (5.55 ± 0.56) , while maintaining a moderate level of task load. This demonstrates that the virtual scenes it constructs effectively stimulate learners' attention and psychological engagement without incurring excessive cognitive burden. This advantage stems from the deep consistency of cultural semantics and the standardized interaction logic, which enables users to gain a stable sense of contextual credibility through behavioral feedback, thereby promoting the formation of embodied cognition. In contrast, StyleGAN3, Diffusion Model, and CycleGAN, while possessing strong visual expressiveness, often exhibit symbol misplacement or behavioral disorganization due to a lack of knowledge guidance, disrupting the user's internal immersion process. These systems exhibit higher mental load on the NASA-TLX scale, indicating that users must invest additional resources to interpret and reconcile perceptual contradictions, weakening the depth of emotional engagement. Notably, HRV trends show that all models induce moderate autonomic nervous system activation, but the KG-GAN model exhibits a more stable LF/HF ratio, reflecting a stimulation rhythm that better aligns with the cognitive rhythm of instruction. Overall, true immersion relies not only on sensory realism but also on the cultural and logical integrity of the content and the semantic coherence of human-computer interaction. Only by internalizing knowledge structures within generation mechanisms can we achieve the synergistic optimization of deep immersion and low external load in educational scenarios.

3 Discussion

The results of this paper demonstrate that KG-GAN significantly outperforms StyleGAN3, diffusion models, and CycleGAN in terms of cultural semantic fidelity, temporal stability, and interaction compliance. Its advantage stems from a deep symbol integration mechanism: unlike existing methods that merely treat cultural elements as surface labels or style cues, KG-GAN embeds a knowledge graph into the entire generation-discrimination game process, making symbolic logic a differentiable constraint. This effectively suppresses temporal misalignment and relational mismatch, particularly excelling in highly structured scenarios such as rituals and architecture.

However, this method still has limitations in abstract symbols (such as "harmony between heaven and man" and "the spirit of ritual and music"). As shown in Figure 4, the confidence score for symbolic relation extraction is only 0.69, reflecting the insufficient ability of current graph modeling to represent metaphorical and non-entity concepts.

Furthermore, while the structural constraints introduced by knowledge embedding improve cultural accuracy, they slightly limit visual diversity—the generated results may sacrifice some artistic freedom under strict adherence to norms, reflecting a trade-off between embedding complexity and visual fidelity.

Finally, KG-GAN exhibits good generalization ability: the knowledge graph can be readily replaced with other cultural systems; however, high-quality performance depends on domain-specific graph construction and manually corrected data, requiring additional optimization in low-resource cultural scenarios. Future work will explore mechanisms for few-shot knowledge injection and cross-cultural transfer.

4 Conclusion

This paper constructs a generative adversarial network framework based on knowledge graph constraints, achieving high-fidelity virtual reconstruction of traditional cultural scenes and optimizing the immersive teaching experience in vocational Chinese language classrooms. By encoding cultural entities and relationships into semantic embedding vectors and integrating them into the generator input and discriminator feedback mechanism, the generated content significantly improves the historical accuracy of clothing, artifacts, spatial layout, and other aspects. A multi-scale training strategy and attention regulation mechanism further enhance the cultural compliance of local details and overall structure. At the dynamic evolution level, knowledge-driven temporal modeling and optical flow continuity control ensure logical coherence and smooth visual transitions in complex scenarios such as ritual processes. A layered interaction design, combined with natural language parsing and behavior mapping, ensures that user operations strictly adhere to ritual norms and avoid contextual conflicts. Subjective and objective evaluations demonstrate that this method outperforms mainstream generative models in terms of cultural semantic fidelity, temporal stability, and immersive perception. Research has confirmed that the generation paradigm that relies solely on data statistical laws is difficult to meet the stringent requirements of educational scenarios for cultural authenticity. Only by deeply coupling structured knowledge with the generation architecture can the organic unity of technological empowerment and humanistic inheritance be achieved, providing a verifiable and scalable technical path for the construction of cultural cognition in an intelligent education environment. This bottleneck stems partly from the limited expressive capacity of translation-based embeddings like TransE for asymmetric and higher-order semantics (e.g., 'symbol' and 'metaphor'). Future research could explore a hybrid representation learning framework that collaborates with symbolic logic reasoning modules.

Notably, this framework's design philosophy shares profound parallels with adaptive/robust control in engineering. The knowledge graph, serving as a "structured prior" akin to a reference model in controllers, maintains semantic stability in output despite input disturbances (e.g., ambiguous texts or cultural ambiguities). The semantic discriminator and attention mechanism together form a "feedback correction loop" that dynamically suppresses cultural drift. This mechanism enables seamless migration to high-fidelity cultural expression scenarios like museum digital exhibitions (e.g., Maya ritual reconstructions) and XR cultural tourism (e.g., ancient Greek theater recreations). Simply replacing the underlying knowledge graph without reconstructing the generative architecture demonstrates excellent generalization and deployment flexibility.

References

[1] Song L. Construction of socialist core values from the perspective of Chinese traditional culture[J]. *International Journal of Frontiers in Sociology*, 2021, 3(12): 69–76. https://doi.org/10.25236/ijfs.2021.031210

[2] Jiang B. Research on the application of Chinese traditional culture teaching in higher vocational education[J]. *Kuram ve Uygulamada Egitim Bilimleri*, 2022, 22(2): 1–13. *(Result score too low)*

[3] Barrett A, Pack A, Guo Y, et al. Technology acceptance model and multi-user virtual reality learning environments for Chinese language education[J]. *Interactive Learning Environments*, 2023, 31(3): 1665–1682. https://doi.org/10.1080/10494820.2020.1855209

[4] Huang W. Digital realization of traditional residential houses in Fujian: application and research on immersive scenes in virtual animation design[J]. *Procedia of Multidisciplinary Research*, 2023, 1(9): 15–15. *(Result score too low)*

[5] Li Z, Lu H, Fu H, et al. Csan: cross-coupled semantic adversarial network for cross-modal retrieval[J]. Artificial Intelligence Review, 2025, 58(5): 1-27. https://doi.org/10.1007/s10462-025-11152-7

[6] Jin S, Fan M, Kadir A. Immersive spring morning in the Han Palace: learning traditional Chinese art via

virtual reality and multi-touch tabletop[J]. *International Journal of Human–Computer Interaction*, 2022, 38(3): 213–226. https://doi.org/10.1080/10447318.2021.1930389

[7] Shadiev R, Wang X, Huang Y M. Cross-cultural learning in virtual reality environment: facilitating cross-cultural understanding, trait emotional intelligence, and sense of presence[J]. *Educational Technology Research and Development*, 2021, 69(5): 2917–2936. https://doi.org/10.1007/s11423-021-10044-1

[8] Wang Y, Wang L, Siau K L. Human-centered interaction in virtual worlds: a new era of generative artificial intelligence and metaverse[J]. *International Journal of Human–Computer Interaction*, 2025, 41(2): 1459–1501. https://doi.org/10.1080/10447318.2024.2316376

[9] Liu Q, Chen H, Crabbe M. Interactive study of multimedia and virtual technology in art education[J]. *International Journal of Emerging Technologies in Learning (iJET)*, 2021, 16(1): 80–93. https://doi.org/10.3991/ijet.v16i01.18227

[10] Li W, Huang H, Solomon T, et al. Synthesizing personalized construction safety training scenarios for VR training[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2022, 28(5): 1993–2002. https://doi.org/10.1109/tvcg.2022.3150510

[11] Qian J. Research on artificial intelligence technology of virtual reality teaching method in digital media art creation[J]. *Journal of Internet Technology*, 2022, 23(1): 125–132. https://doi.org/10.53106/160792642022012301013

[12] Pirker J, Dengel A. The potential of 360 virtual reality videos and real VR for education—a literature review[J]. *IEEE Computer Graphics and Applications*, 2021, 41(4): 76–89. https://doi.org/10.1109/mcg.2021.3067999

[13] Hammady R, Ma M, Al-Kalha Z, et al. A framework for constructing and evaluating the role of MR as a holographic virtual guide in museums[J]. *Virtual Reality*, 2021, 25(4): 895–918. https://doi.org/10.1007/s10055-020-00497-9

[14] Zhong H, Wang L, Zhang H. The application of virtual reality technology in the digital preservation of cultural heritage[J]. *Computer Science and Information Systems*, 2021, 18(2): 535–551. https://doi.org/10.2298/csis200208009z

[15] Karuzaki E, Partarakis N, Patsiouras N, et al. Realistic virtual humans for cultural heritage applications[J]. *Heritage*, 2021, 4(4): 4148–4171. https://doi.org/10.3390/heritage4040228

[16] Dong X, Mao Z, Sun Y, et al. Short-term wind power scenario generation based on conditional latent diffusion models[J]. *IEEE Transactions on Sustainable Energy*, 2023, 15(2): 1074–1085. https://doi.org/10.1109/tste.2023.3327497

[17] Peng J, Qiu R L J, Wynne J F, et al. CBCT-based synthetic CT image generation using conditional denoising diffusion probabilistic model[J]. *Medical Physics*, 2024, 51(3): 1847–1859. https://doi.org/10.1002/mp.16704

[18] Zheng L. Artistic style image migration model based on cycle-consistent generative adversarial networks[J]. *International Journal of Information and Communication Technology*, 2025, 26(16): 53–68. https://doi.org/10.1504/ijict.2025.146377

[19] Yuan R, Wang B, Sun Y, et al. Conditional style-based generative adversarial networks for renewable scenario generation[J]. *IEEE Transactions on Power Systems*, 2022, 38(2): 1281–1296. https://doi.org/10.1109/tpwrs.2022.3170992

[20] Kang M, Zhu R, Chen D, et al. A cross-modal generative adversarial network for scenarios generation of renewable energy[J]. *IEEE Transactions on Power Systems*, 2023, 39(2): 2630–2640. https://doi.org/10.1109/tpwrs.2023.3277698

[21] Gan M, Wang C. Esophageal optical coherence tomography image synthesis using an adversarially learned variational autoencoder[J]. *Biomedical Optics Express*, 2022, 13(3): 1188–1201. https://doi.org/10.1364/boe.449796

[22] Wang Y, Ma X, Wang J, et al. Robust AUV visual loop-closure detection based on variational autoencoder network[J]. *IEEE Transactions on Industrial Informatics*, 2022, 18(12): 8829–8838. https://doi.org/10.1109/tii.2022.3145860

[23] Zhang P, Wang C, Kumar N, et al. Dynamic virtual network embedding algorithm based on graph convolution neural network and reinforcement learning[J]. *IEEE Internet of Things Journal*, 2021, 9(12): 9389–9398. https://doi.org/10.1109/jiot.2021.3095094

[24] Suárez-Varela J, Almasan P, Ferriol-Galmés M, et al. Graph neural networks for communication networks: context, use cases and opportunities[J]. *IEEE Network*, 2022, 37(3): 146–153. https://doi.org/10.1109/mnet.123.2100773

[25] Lim S H, Park J W. Hierarchical control strategy for effective virtual frequency responses of multiple WPPs[J]. *IEEE Transactions on Power Systems*, 2023, 39(1): 576–586. https://doi.org/10.1109/tpwrs.2023.3269052

[26] Zeng X, Peng H, Su D, et al. Hierarchical decision making based on structural information principles[J]. Journal of Machine Learning Research, 2025, 26(182): 1-55. http://jmlr.org/papers/vxx/xx-xxxx.html

[27] Wang J, Wu Q M J, Pourpanah F. An attentive-based generative model for medical image synthesis[J]. *International Journal of Machine Learning and Cybernetics*, 2023, 14(11): 3897–3910. https://doi.org/10.1007/s13042-023-01871-0

[28] Liu X, Pan J, Li X, et al. Attention based cross-domain synthesis and segmentation from unpaired medical images[J]. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023, 8(1): 917–929. https://doi.org/10.1109/tetci.2023.3296499