IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 3, June, 2026, pp. 1307-1321
Journal homepage: http://iieta.org/journals/ts

Intelligent Modeling of University Cultural Content via Cross-Modal Image Processing

Xiaoqian Zhang[1*], Yi Tian[2]

[1] Education and Teaching Research Center, Shangluo University, Shangluo 726000, China
[2] Public Big Data Research Center of Shangluo, Shangluo University, Shangluo 726000, China

Corresponding Author Email: zxqian@slxy.edu.cn

https://doi.org/10.18280/ts.430318

Received: 22 December 2025
Revised: 20 May 2026
Accepted: 29 May 2026
Available online: 30 June 2026

Keywords:
cross-modal image processing, cultural semantic preservation, orthogonal feature decoupling, knowledge graph contrastive learning, conditional diffusion models, university cultural digitization

ABSTRACT

The convergence of digital humanities and visual intelligence technologies has established novel pathways for the digital preservation, regeneration, and dissemination of university cultural resources, with cross-modal image processing emerging as a fundamental enabling technique for intelligent modeling of campus culture. Existing cross-modal modeling and image generation methods, however, are inadequately adapted to the hierarchical semantic characteristics inherent to university cultural scenarios, and persistent challenges impede the accurate modeling and faithful regeneration of cultural visual content. Therefore, a semantic-preserving cross-modal intelligent modeling framework specifically oriented toward university cultural scenes was proposed. In this framework, visual feature extraction was optimized via a cultural saliency attention mechanism, and orthogonal constraint-based decoupling networks were employed to achieve independent subspace representations of low-level visual morphological information and high-level cultural semantic information, precisely purifying cultural features. A multi-level contrastive learning mechanism, structured upon a constructed university cultural knowledge graph, was further established to progressively realize fine-grained cross-modal semantic alignment across sample matching, entity association, and topological relationship dimensions. On this basis, the unified semantic representation was incorporated as a prior condition into the reverse denoising process of a diffusion model, and a customized semantic constraint loss was formulated to suppress semantic drift in generation tasks, ensuring cultural consistency in visual generation and enhancement outcomes. Systematic experimental validation was conducted using a self-constructed university cultural multimodal dataset and a publicly available cultural heritage dataset. Results demonstrated that the proposed method significantly outperformed existing state-of-the-art algorithms across core tasks, including cross-modal retrieval, image enhancement, and controllable cultural image generation. This study effectively overcomes the technical bottlenecks of cross-modal modeling in culture-specific vertical scenarios, provides a specialized solution for the digital inheritance of university culture, and offers a viable technical paradigm for visual intelligence modeling research within the digital humanities domain.

1. INTRODUCTION

The rapid iteration of digital humanities technologies has propelled traditional cultural heritage into a comprehensive digital transformation phase. The campus architecture, historical imagery, campus activities, and humanistic spirits accumulated by higher education institutions collectively constitute a distinctive university cultural system [1, 2], which serves as a significant carrier for regional culture and the transmission of higher education cultural heritage. In contrast to general visual cultural resources, university cultural images encapsulate hierarchical historical connotations and humanistic values [3], embodying not only fundamental visual morphological features but also culture-specific symbolic semantics and historical temporal information. The digital faithful preservation, intelligent restoration, and innovative regeneration [4] of such resources represent core research directions within contemporary cultural digitization initiatives. At present, technical demands for historical campus image quality enhancement [5], controllable generation of cultural scenes [6], and cross-modal intelligent retrieval of cultural semantics [7] continue to escalate. Conventional image processing methodologies, however, are capable only of optimization at the basic visual level and cannot achieve precise preservation or in-depth parsing of cultural semantics [8, 9]. With the progressive maturation of cross-modal vision-language modeling techniques, the deep integration of structured domain knowledge with visual generation and feature alignment methodologies [10] has emerged as a prevailing developmental trend in vertical-scenario intelligent modeling. By constraining the visual modeling process through a dedicated cultural knowledge framework, the limitations of generic models in cultural semantic understanding can be effectively overcome, thereby achieving a dual unification of visual quality optimization and cultural semantic fidelity. This holds substantial theoretical value and

engineering practical significance for advancing the digital preservation, intelligent dissemination, and innovative application of university culture.

Although cross-modal representation learning and image generation technologies have been extensively applied in the cultural image processing domain, significant technical deficiencies persist in existing methods when deployed in university culture-specific scenarios, rendering them inadequate for meeting the application demands of high-precision cultural intelligent modeling. Within current visual encoding systems, image features are extracted via a globally undifferentiated approach, without distinguishing between low-level visual morphological information and high-level cultural semantic information. This results in a high degree of entanglement between the two feature types within the representational space [11, 12]. Irrelevant background textures and visual noise continuously dilute core cultural features, thereby preventing models from accurately capturing the culture-specific semantic information of campus architecture, cultural symbols, and historical scenes, and substantially constraining the upper performance bounds of cross-modal matching and intelligent generation tasks [13, 14]. At the level of cross-modal semantic alignment, mainstream image-text contrastive learning methods rely solely on sample-level global pairing relationships for model training, lacking structured and hierarchical semantic constraint mechanisms. Such coarse-grained alignment paradigms are capable only of holistic semantic matching between images and texts, and are unable to accommodate the entity units and relational systems inherent to university culture. Consequently, precise alignment of fine-grained cultural entities and semantic associations cannot be achieved, and refined cultural retrieval and semantic parsing tasks cannot be effectively supported. In the domain of image generation and restoration, existing diffusion-based generative models are optimized with a primary focus on appearance-oriented metrics, such as visual clarity and textural details, without constructing semantic constraint mechanisms targeted at cultural content [15, 16]. During the restoration of historical photographs and the regeneration of campus cultural scenes, such models are highly prone to cultural symbol morphological distortion, historical scene style deviation, and humanistic semantic loss, thereby failing to guarantee the cultural authenticity and integrity of generated content [17, 18]. Overall, a full-process cross-modal image processing system adapted to university cultural scenarios has yet to be established within the academic community, and the associated technical gaps constrain the practical deployment of intelligent digital modeling for university culture.

From the perspective of university cultural governance and educational practice, the digitization of university cultural resources is not merely a matter of preserving historical materials and improving image quality; rather, it should serve the development of quality cultural construction within higher education institutions and the enhancement of cultural education systems. By transforming campus architecture, historical photographic archives, cultural symbols, and spiritual lineages into identifiable, retrievable, and generatable intelligent cultural content, the transition of university culture from static exhibition to dynamic dissemination, from experience-based accumulation to data-driven modeling, and from singular resource storage to the reconstruction of educational scenarios can be facilitated. Such a transformation provides technical support for higher education institutions in establishing a digital humanities ecosystem characterized by continuous improvement awareness, cultural identity foundations, and value-guiding functions.

To address the aforementioned technical bottlenecks, a full-process cross-modal intelligent modeling framework oriented toward university cultural scenes is constructed in this study, providing an integrated solution spanning feature representation, semantic alignment, and image generation. A cultural saliency-guided orthogonal feature decoupling network is established, enabling the separation and purification of visual morphology and cultural semantic features, and thereby fundamentally resolving the representation challenge caused by cultural feature entanglement. In conjunction with a university cultural knowledge graph, a multi-level contrastive learning strategy is designed, and a refined structured cross-modal alignment mechanism is established, effectively enhancing the matching accuracy of fine-grained cultural semantics. The diffusion-based generative model is optimized through semantic constraint incorporation, and a cultural semantic-preserving loss function is formulated to suppress semantic drift during image restoration and generation processes. Furthermore, an image-text knowledge triple university cultural multimodal dataset is constructed, and a unified general semantic space is established, which can provide both data and technical support for multiple downstream cultural vision tasks.

The overall research is organized according to a logical progression encompassing problem analysis, method construction, experimental validation, and conclusion with future directions. Chapter 2 elaborates on the overall architecture of the proposed framework, the dataset construction rules, and the technical principles of each core module. Chapter 3 designs multi-dimensional comparative experiments, ablation studies, and generalization experiments, combining quantitative and qualitative analyses to validate method performance. Chapter 4 provides an in-depth analysis of the model's operational mechanisms, objectively examining the application boundaries and practical value of the proposed approach. Chapter 5 synthesizes the research findings, summarizes the innovative advantages, and outlines directions for future investigation.

2. CROSS-MODAL INTELLIGENT MODELING FRAMEWORK FOR CULTURAL SEMANTIC PRESERVATION

2.1 Overall framework architecture

A cross-modal intelligent modeling framework for cultural semantic preservation is constructed in this study, establishing an end-to-end integrated optimization system tailored for image understanding, enhancement, and generation tasks within university cultural scenes. The overall framework follows a progressive technical logic of feature purification, semantic alignment, and controllable generation, with full-process supervised training enabled by multimodal triple data. Within this framework, multi-scale visual features are first extracted via a culture-aware visual encoding module, and the separation and purification of low-level visual morphological information from high-level cultural semantic information are accomplished through an orthogonal decoupling mechanism, thereby eliminating the interference of irrelevant visual noise on cultural representations. The decoupled high-quality

cultural visual features, in conjunction with textual semantic features and structured knowledge graph embeddings, are then processed through a hierarchical contrastive learning module to achieve multi-granularity cross-modal semantic alignment, from which standardized cultural representations endowed with structured semantic logic are produced. These representations are further incorporated as conditional priors into the reverse denoising process of the diffusion model, enabling semantic constraint optimization for cultural image enhancement and controllable generation. Finally, all modal features are mapped into a unified space via a semantic bridging layer, constructing a general semantic representation space that supports multiple downstream vision tasks, including cross-modal retrieval and semantic reasoning.

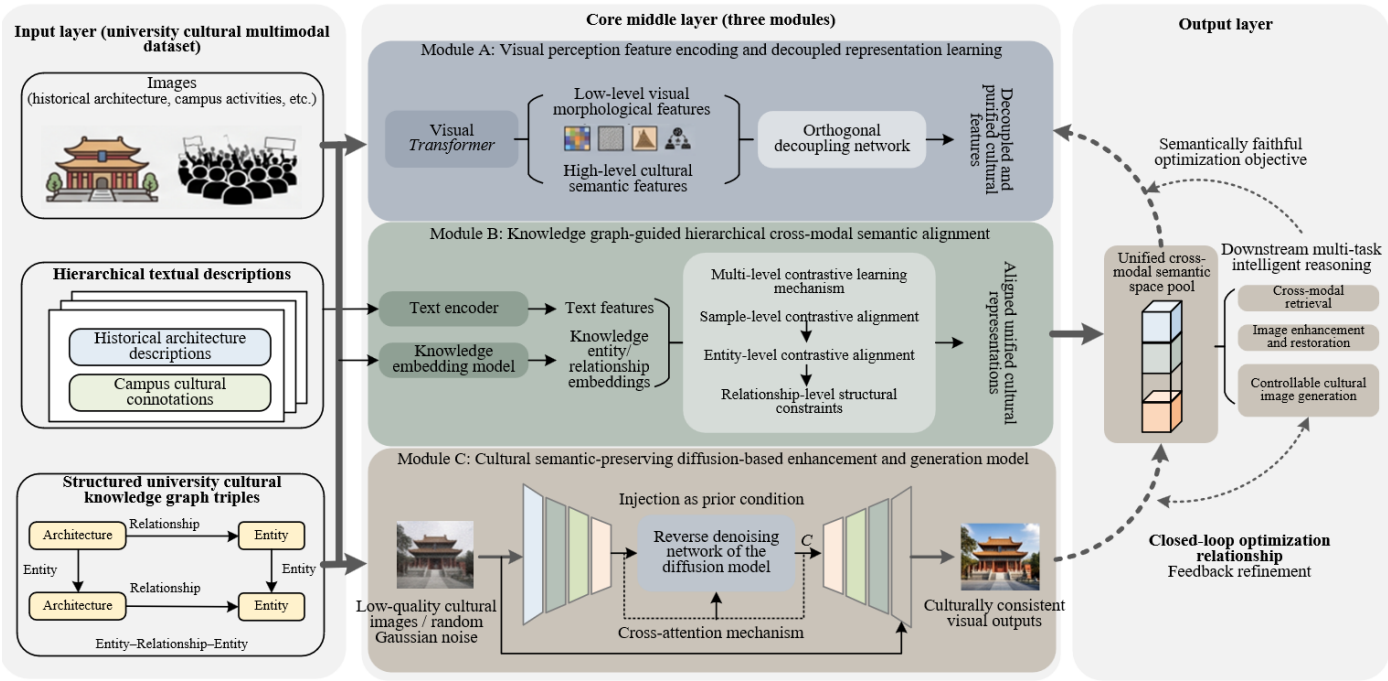

Figure 1. Overall framework of the semantic-preserving cross-modal intelligent modeling system for university cultural scenes

The overall architecture of the semantic-preserving cross-modal intelligent modeling framework for university cultural scenes is illustrated in Figure 1. The core innovative modules of the entire framework form a deeply coupled closed-loop optimization system, which differs fundamentally from the modularly independent iterative design paradigm of conventional methods. The feature decoupling module provides purified cultural semantic features for cross-modal alignment tasks, substantially reducing matching deviations caused by morphological noise. The knowledge graph-driven multi-level alignment mechanism supplies precise structured semantic anchors for image generation tasks, effectively addressing the issue of cultural semantic drift during the generation process. Concurrently, the semantically faithful optimization objective of the generation task enables the backpropagation of gradients, which continuously refines the front-end feature extraction and semantic alignment processes, thereby realizing bidirectional empowerment among representation learning, semantic matching, and visual generation. Through this systemic integration, the hierarchical and high-fidelity intelligent modeling requirements of university culture are comprehensively accommodated.

2.2 University cultural multimodal dataset construction

To support cross-modal modeling training for cultural semantic preservation and multi-task performance validation, a multimodal triple dataset oriented toward university cultural scenes is constructed in this study, establishing a standardized data system centered on visual images, with textual descriptions serving as fine-grained supervision and a knowledge graph providing structured semantic constraints, thereby achieving precise matching and logical association among the three modal data types. The visual samples within the dataset encompass four major campus cultural scene categories: historical architecture, cultural landscapes, cultural relic collections, and campus activities. The collection process incorporates sample variations across different lighting environments, shooting angles, and seasonal conditions, fully reflecting the complex distribution characteristics of real-world campus imagery. The dataset concurrently includes a substantial volume of low-resolution historical photographs and scanned materials, accommodating the training requirements of both general cross-modal understanding tasks and degraded image enhancement and restoration tasks. Textual annotation is implemented through a hierarchical semantic description system, with fine-grained annotations performed across three dimensions—local visual element features, global scene content, and deep cultural connotations—thereby producing multi-scale textual supervision information that aligns with the training logic of hierarchical cross-modal alignment in the model. A dedicated university cultural knowledge graph is constructed by drawing upon authoritative institutional historical documents and official materials, within which multiple categories of cultural entities and semantic relationships are formally defined, and the intrinsic topological regularities of campus cultural elements are represented in the form of structured triples, providing stable domain prior support for fine-grained cross-modal semantic alignment.

The dataset is randomly partitioned into training, validation, and test sets according to a fixed ratio, with the partitioning

process rigorously ensuring balanced distribution across cultural scene categories and sample quality levels, thereby preventing data bias from interfering with model training. All image annotations, textual descriptions, and knowledge triple contents are subjected to a two-round cross-verification mechanism, through which manual annotation errors are minimized to the greatest extent possible, and the standardization and reliability of the dataset are ensured. This dataset addresses the gap in standardized multimodal resources within the university cultural domain and can provide comprehensive data support for various vertical vision tasks, including cross-modal retrieval, image restoration and enhancement, and controllable cultural content generation.

2.3 Culture-aware visual encoding and decoupled representation learning

A culture-aware encoding network is constructed in this study based on the visual Transformer architecture. To address the deficiencies of general vision models in campus cultural image representation—specifically, insufficient focus on key regions and the susceptibility of cultural semantics to suppression by background noise—cultural saliency priors are introduced to enable adaptive optimization of the attention mechanism. The architecture of the culture-aware visual encoding and orthogonal feature decoupling network is illustrated in Figure 2.

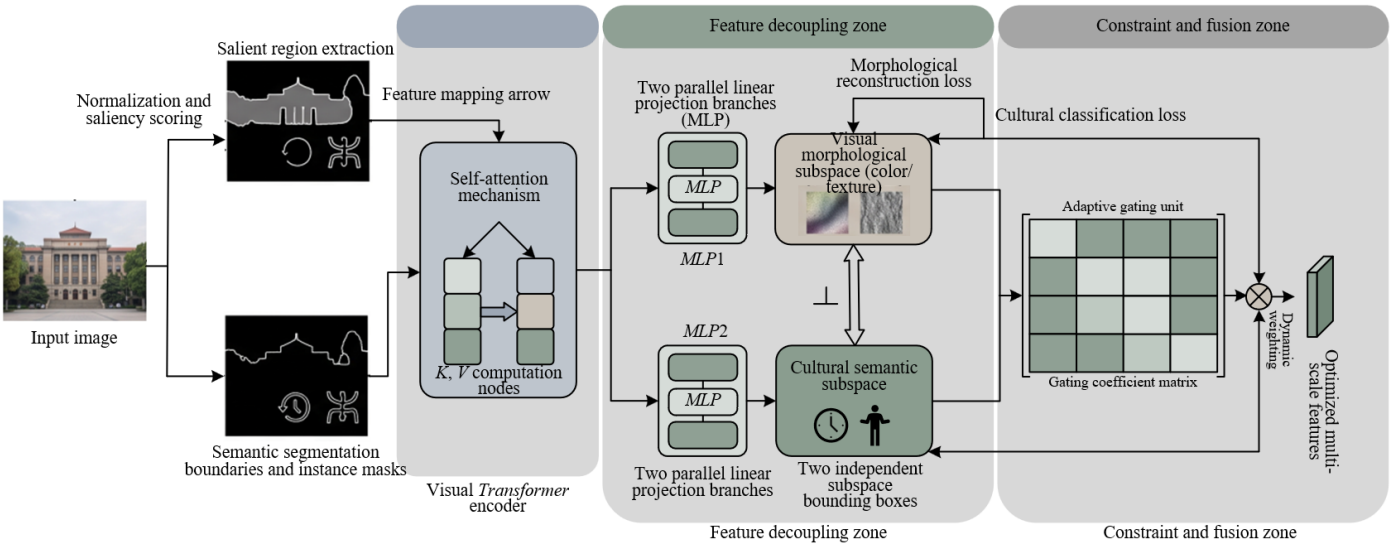

Figure 2. Architecture of the culture-aware visual encoding and orthogonal feature decoupling network

Within this module, the input image is partitioned into uniformly sized image patches, from which foundational image embedding features are generated through linear projection and positional encoding, thereby completing the structured representation of low-level visual information. To guide the model in prioritizing target regions that carry core cultural connotations, a detection model fine-tuned on the campus cultural dataset is employed to generate a global saliency heatmap, through which the cultural semantic density of each image region is quantified. The heatmap is then downsampled to match the image patch granularity, from which normalized saliency scores are obtained. These scores are incorporated as bias terms into the self-attention computation process, and the optimized attention computation formula can be expressed as:

$$Attn_{i,j}=softmax\left(\frac{(W_Q e_i)(W_K e_j)^T}{\sqrt{d_k}}+\lambda\cdot(s_i+s_j)\right) \quad (1)$$

where, W_Q and W_K denote the query and key projection matrices, respectively; d_k is the attention head dimension; s_i and s_j are the saliency scores of the corresponding image patches; and λ is used to regulate the constraint strength of the saliency prior. Through this mechanism, the feature responses of cultural entity regions are enhanced, while ineffective background interference is suppressed, on the basis of original semantic relevance modeling. Ultimately, a multi-scale Transformer feature set is produced, encompassing hierarchical semantic information ranging from local textural details to global cultural scenes, thereby providing comprehensive preliminary representational support for the subsequent feature decoupling task.

To address the challenge of high coupling between low-level visual morphology and high-level cultural semantics in university cultural images, an orthogonal constraint-based cultural feature decoupling network is constructed in this study, through which the explicit separation and independent representation of these two heterogeneous information types are achieved. Two parameter-independent linear projection branches are established within the network, and the multi-layer visual features output by the encoder are mapped into subspaces via these branches, from which feature matrices specialized for visual morphological representation and cultural semantic representation are respectively generated. The mapping process can be formulated as follows:

$$Z_{shape}^{(l)}=\sigma(W_{shape}F_v^{(l)}+b_{shape}) \quad (2)$$

$$Z_{cult}^{(l)}=\sigma(W_{cult}F_v^{(l)}+b_{cult}) \quad (3)$$

where, W_{shape}, W_{cult}, b_{shape}, and b_{cult} denote the weights and biases of the two projection branches, respectively; and σ represents the rectified linear unit activation function. The morphological subspace features preserve physical visual information, including color, texture, and spatial composition, while the cultural subspace features are dedicated to storing humanistic semantic information, including historical connotations and symbolic meanings. To prevent re-

entanglement of the two feature types, an orthogonality loss based on the Frobenius norm is introduced to constrain subspace independence:

$$L_{orth}=\sum_{l}\left\|(Z_{shape}^{(l)})^{T}Z_{cult}^{(l)}\right\|_{F}^{2} \tag{4}$$

Simultaneously, a morphological reconstruction loss and a cultural classification loss are introduced to constrain the representational effectiveness of the two subspaces, respectively, ensuring that the morphological features can fully reconstruct the visual content of the image and that the cultural features possess precise semantic discriminative capability. The overall loss function of the module is optimized through the weighted integration of the three constraints:

$$L_{disentangle}=L_{orth}+\mu_1 L_{recon}+\mu_2 L_{cult\text{-}cls} \tag{5}$$

where, μ_1 and μ_2 are hyperparameters used to balance the optimization weights of the different loss terms, ensuring that the decoupling process accommodates both information completeness and semantic specificity.

Following feature decoupling, an adaptive gating feature interaction unit is designed in this study, through which the dynamic fusion of visual morphological features and cultural semantic features is achieved, accommodating the representational requirements of diverse campus cultural scenes. Within this unit, the two-layer decoupled features are first concatenated along the channel dimension, and a single-layer perceptron with a Sigmoid activation function is employed to generate pixel-wise adaptive gating coefficients, through which the dynamic assessment of scene feature weights is accomplished:

$$g^{(l)}=sigmoid(W_g[Z_{shape}^{(l)};Z_{cult}^{(l)}]+b_g) \tag{6}$$

where, W_g and b_g denote the learnable parameters of the gating unit, and the gating coefficients take values in the range of [0, 1]. Based on the obtained weights, the weighted fusion of the dual features is performed, and the optimized multi-scale features are finally output as follows:

$$F_{fuse}^{(l)}=g^{(l)}\odot Z_{shape}^{(l)}+(1\text{-}g^{(l)})\odot Z_{cult}^{(l)} \tag{7}$$

where, $\odot$ denotes element-wise multiplication. Through this adaptive mechanism, the feature proportions can be dynamically adjusted according to the content characteristics of the image: for texture-rich modern campus scenes, the preservation of visual morphological features is prioritized, whereas for culturally rich scenes such as historical architecture and cultural relic imagery, the weighting of semantic features is enhanced. This enables precise matching between feature representations and scene attributes.

The complete encoding and decoupling representation system optimizes the representation quality of cultural images from the feature extraction source. Through the saliency attention mechanism, key cultural features are purified; through orthogonal constraints, the precise decoupling of heterogeneous information is achieved; and through the gated fusion mechanism, adaptive feature optimization is accomplished. The final output fused features possess both complete visual structural information and highly discriminative cultural semantic information, fundamentally resolving the issues of semantically mixed representations and weakened cultural features inherent in conventional visual representations. A high-quality visual representation foundation is thereby provided for subsequent cross-modal semantic alignment and culturally faithful image generation tasks.

2.4 Knowledge graph-guided hierarchical contrastive alignment mechanism

To overcome the limitations of conventional image-text contrastive learning, which relies solely on global sample-pair supervision and cannot adequately accommodate the hierarchical semantic system of university culture, a structured university cultural knowledge graph is introduced in this study, through which a multi-granularity semantic constraint system is constructed, and fine-grained alignment of cross-modal features is achieved via progressive hierarchical contrastive learning. The schematic diagram of the knowledge graph-guided hierarchical cross-modal contrastive alignment mechanism is illustrated in Figure 3. The university cultural knowledge graph is defined as a structured set of triples G=(E,R,T), where E, R, and T denote the set of cultural entities, the set of semantic relations, and the set of knowledge triples, respectively. The RotatE algorithm is employed to accomplish the vectorized representation of the knowledge graph, through which discrete cultural entities and semantic relations are mapped into a continuous vector space of dimension d_k, and entity embedding features (e) and relation embedding features (r) are generated as output. The embedding dimension is maintained in strict consistency with the representation dimensions of the visual and textual modalities, thereby enabling seamless integration across multimodal spaces. The knowledge embedding process is trained via a standard triple ranking loss and is jointly optimized with the cross-modal contrastive task. Through entity-level feature mapping, deep coupling between the knowledge semantic space and the visual-textual space is achieved, providing stable domain prior support for multi-level fine-grained semantic alignment.

A three-level progressive contrastive learning paradigm—encompassing sample, entity, and relational levels—is constructed in this study, and a unified optimization objective is formulated through the weighted integration of multi-granularity constraint losses, through which cross-modal semantic features are progressively converged from coarse to fine granularity. The overall loss function is expressed as:

$$L_{CL}=\alpha L_{ins}+\beta L_{entity}+\gamma L_{rel} \tag{8}$$

where, α, β, and γ denote the weight hyperparameters of the three loss terms, used to balance the optimization strengths of global sample matching, entity-level semantic alignment, and topological relationship constraints. Sample-level contrastive learning serves as the fundamental alignment unit, through which the matching optimization of global image and text features is accomplished via the information noise-contrastive estimation loss:

$$L_{ins}=-\frac{1}{B}\sum_{i=1}^{B}\log\frac{exp(sim(f_v(I_i),f_t(T_i))/\tau)}{\sum_{j=1}^{B}exp\,(sim(f_v(I_i),f_t(T_j))/\tau)} \tag{9}$$

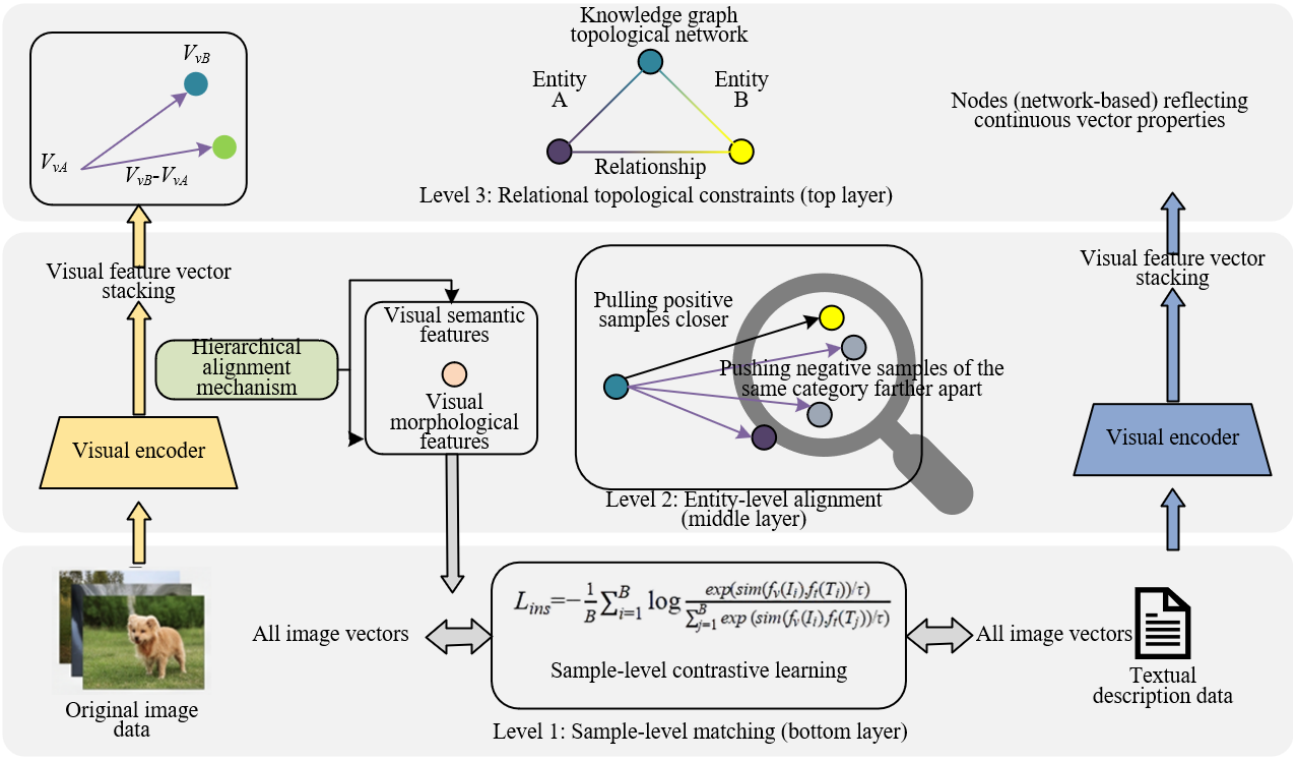

Figure 3. Schematic diagram of the knowledge graph-guided hierarchical cross-modal contrastive alignment mechanism

where, B denotes the training batch size; f_v and f_t are the feature mapping functions of the visual encoder and the text encoder, respectively; *sim* denotes the cosine similarity between vectors; and τ is the temperature coefficient used to regulate the dispersion of the feature distribution. On this basis, an entity-level contrastive constraint is further introduced, through which the conventional global coarse-grained matching paradigm is abandoned. Standardized cultural entities from the knowledge graph are adopted as semantic anchors, and a fine-grained contrastive loss is constructed via a same-category heterogeneous negative sample sampling strategy:

$$L_{entity}=-\frac{1}{B}\sum_{i=1}^{B}\log\frac{exp(sim(f_v(I_i),e_i)/\tau)}{\sum_{e'\in N(e_i)}exp\,(sim(f_v(I_i),e')/\tau)} \tag{10}$$

Through this strategy, negative sample sets are constructed from same-source broad-category cultural entities, through which the model's discriminative capability for similar cultural scenes and same-category cultural symbols is effectively enhanced, and visual and textual features are precisely mapped to culture-specific entity semantics, thereby circumventing the matching ambiguity issues caused by superficial textual descriptions.

To further preserve the structured associative characteristics of the cultural system, a relationship-level contrastive loss is introduced to constrain the topological consistency of the embedding space, addressing the deficiency that isolated entity alignment cannot represent semantic associations. This constraint is based on the established entity association rules of the knowledge graph, through which the model is forced to learn the fixed semantic offset regularities among entities. The loss function is defined as:

$$L_{rel}=\sum_{(e_i,r,e_j)\in T}\|(f_v(I_i)-f_v(I_j))-r\|_2^2 \tag{11}$$

This formulation constrains the visual feature difference between any two associated entities to remain consistent with the predefined relation embedding vector from the knowledge graph, thereby enabling the cross-modal semantic space to fully inherit the structured logic of the knowledge graph, and achieving an upgrade from isolated feature matching to systematic semantic modeling. The three-tier contrastive constraints form a progressively layered optimization system: the basic sample-level loss ensures the accuracy of global image-text matching, the entity-level loss strengthens the discriminative capability of fine-grained cultural semantics, and the relationship-level loss solidifies the topological structure of cultural knowledge. Ultimately, a cross-modal semantic space possessing both global matching precision and local structural integrity is constructed, providing a highly consistent representational foundation for subsequent culturally faithful semantic preservation generation tasks.

2.5 Cultural semantic-preserving diffusion-based enhancement and generation model

To address the issue of cultural information drift in conventional diffusion models during cultural image restoration and generation—caused by excessive optimization of visual textures without constraints on semantic consistency—a cultural semantic-preserving diffusion-based enhancement and generation model is constructed in this study. The structured cultural semantic features output by the preceding modules are incorporated as prior conditions into the diffusion inference process, through which the collaborative optimization of visual quality enhancement and cultural semantic fidelity is achieved. The schematic diagram of the cultural semantic-preserving diffusion-based enhancement and generation model is illustrated in Figure 4. The forward diffusion process of the model follows the standard Gaussian noise iterative superposition paradigm, in which progressive noise perturbations are applied to the

original clear image, and a continuous noise degradation mapping relationship is established. This process can be formulated as:

$$I_t=\sqrt{\bar{\alpha}_t}I_0+\sqrt{1-\bar{\alpha}_t}\epsilon,\epsilon\sim N(0,I) \tag{12}$$

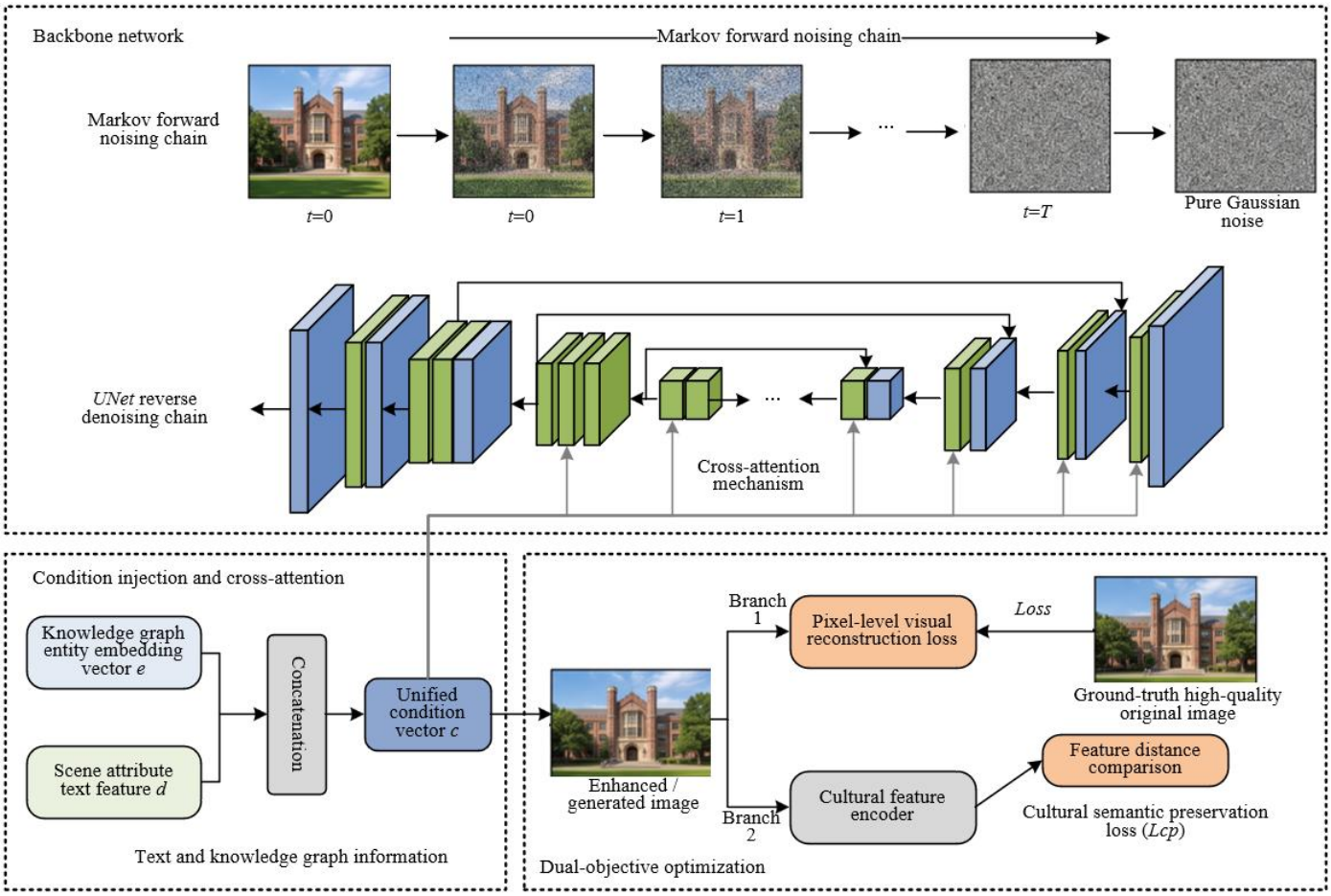

Figure 4. Schematic diagram of the cultural semantic-preserving diffusion-based enhancement and generation model

where, I_0 denotes the original input image; I_t denotes the noisy image at step t ; α_t is the predefined noise scheduling coefficient; $\bar{\alpha}_t$ is the cumulative product of the coefficients from the first t steps; and ϵ is a Gaussian noise vector following a standard normal distribution. The core improvement of this study is concentrated in the diffusion reverse denoising stage, where cultural semantic conditional guidance is introduced on the basis of the original unconstrained denoising network, and a conditional probability generation model is constructed:

$$p_\theta(I_{t-1}|I_t,c)=N(I_{t-1};\mu_\theta(I_t,t,c),\Sigma_\theta(I_t,t,c)) \tag{13}$$

where, μ_θ and Σ_θ denote the learnable mean and variance prediction networks, respectively; and c is the standardized cultural semantic condition vector, which is obtained through the fusion of knowledge graph entity embeddings and the aligned textual semantic features. This condition vector is embedded into the multi-layer feature layers of the UNet network via a cross-attention mechanism, through which global semantic constraints are imposed on each denoising iteration step, ensuring that the denoising optimization process of the model remains consistently aligned with the inherent semantic logic of university culture.

For the restoration and super-resolution enhancement tasks of low-quality historical campus images, a cultural semantic preservation constraint is incorporated into the standard diffusion noise prediction loss, and a multi-objective joint optimization paradigm is constructed, through which cultural feature distortion during the restoration process is fundamentally prevented at the semantic level. The model takes the low-quality image I_{low} as input and the corresponding high-quality original image I_{high} as the visual supervisory benchmark, through which the image quality restoration mapping is accomplished. To ensure that the cultural connotations of the enhanced image remain completely consistent with those of the original image, the cultural feature encoder trained in Section 2.3 is employed to extract high-level semantic features, with the parameters of this encoder frozen during the diffusion model training process, thereby ensuring the stability and discriminability of the cultural representations. The corresponding cultural semantic preservation loss is defined as:

$$L_{cult\text{-}preserve}=\|Enc_{cult}(\hat{I})-Enc_{cult}(I_{high})\|_2^2 \tag{14}$$

where, $\hat{I}$ denotes the enhanced image output by the model, and Enc_{cult} represents the cultural semantic encoding function. Through the minimization of the cultural feature distance between the enhanced image and the ground-truth high-quality image, the model is forced to preserve the complete cultural symbol morphology, historical scene style, and deep

humanistic semantics, while simultaneously optimizing visual metrics such as textural details and sharpness. The deficiencies of conventional image enhancement methods—including cultural element deformation and semantic loss—are thereby effectively addressed.

Leveraging the unified diffusion model backbone architecture, knowledge-guided controllable cultural image generation is further implemented, through which the precise generation and style replication of arbitrary campus cultural scenes is accomplished. To integrate structured knowledge information with fine-grained textual attribute information, the cultural entity embeddings and scene attribute text features are concatenated along the channel dimension, and a multi-layer perceptron is employed to accomplish feature dimension fusion and semantic calibration, from which a unified condition vector adapted for diffusion model input is generated:

$$c=MLP([e;d]) \tag{15}$$

where, e denotes the knowledge graph cultural entity embedding, and d denotes the scene attribute text features. Using the fused semantic vector as the constraint condition, iterative denoising sampling is performed on the random Gaussian noise variable, and a generated image conforming to cultural semantic specifications is finally output. The generation process can be expressed as:

$$I_{gen}=DDPM_{\theta}(z,c),z\sim N(0,I) \tag{16}$$

To further enhance the historical authenticity and stylistic consistency of the generated content, a cultural style consistency loss is introduced. Through the statistical analysis of the color distributions, textural features, and period-specific style regularities of real historical campus images, the visual style distribution of the generated images is constrained to remain consistent with that of authentic cultural scenes, thereby effectively avoiding issues such as anachronistic temporal features and culturally incongruent elements in the generation results.

This module fully inherits the outputs of the preceding feature decoupling and hierarchical semantic alignment stages, with the purified high-precision structured cultural semantics incorporated as constraint priors throughout the entire training and inference process of the diffusion model. In contrast to the optimization logic of generic diffusion models, which focus exclusively on visual pixel fitting, this paradigm achieves the bidirectional constraint of visual appearance optimization and cultural semantic fidelity, and establishes an image enhancement and generation mechanism adapted to university cultural scenes. Both the visual quality and detailed fidelity of the output images are thereby ensured, while the complete preservation of cultural symbols, historical styles, and humanistic connotations is simultaneously achieved, providing reliable technical support for the digital regeneration of university culture.

2.6 Unified cross-modal semantic space construction and inference

To eliminate the modal heterogeneity biases among visual, textual, knowledge graph, and diffusion-generated features, and to achieve the fusion, reuse, and general-purpose inference of multi-source cultural information, a cross-modal semantic bridging layer is constructed in this study, through which the four types of heterogeneous features are uniformly mapped into a shared semantic space S of dimension d. The semantic bridging layer is equipped with independent linear projection matrices and layer normalization units for each modality, through which feature dimension unification and distribution standardization are accomplished while preserving the inherent semantic characteristics of each modality, and the issues of spatial misalignment and scale inconsistency across different representation systems are fundamentally resolved at the structural level. To achieve deep fusion and alignment of multimodal features, a joint optimization loss function is constructed from two dimensions: global distribution matching and local semantic constraints. The overall unification loss is expressed as:

$$L_{unify}=\sum_{m,n\in\{v,t,k,d\}} KL\,(p_m\|p_n)+\sum_i \| s_i^{(v)}-s_i^{(t)}\|_2^2 \tag{17}$$

where, v, t, k, and d denote the visual, textual, knowledge graph, and diffusion-generated modalities, respectively; p_m and p_n represent the feature probability distributions of different modalities; and the Kullback–Leibler divergence is employed to minimize the overall distribution discrepancies across modalities, thereby achieving global spatial distribution alignment. $s_i^{(v)}$ and $s_i^{(t)}$ denote the sample-paired unified semantic vectors of the visual and textual modalities, and the cross-modal representation distances of same-source samples are minimized through L2 distance constraints, thereby ensuring the precision of fine-grained semantic matching.

Based on the optimized unified semantic space, all modal inputs can be transformed into standardized, same-dimensional cultural semantic vectors, through which a general-purpose representation system adapted to university cultural scenes is constructed. This space fully inherits the output advantages of the preceding feature decoupling, hierarchical semantic alignment, and semantically constrained generation modules, achieving deep fusion of the representational outputs from multiple modules. Without requiring parameter fine-tuning for individual downstream tasks, stable support is provided for various intelligent inference tasks, including cross-modal retrieval, image caption generation, visual question answering, and controllable cultural image generation. Through this design, the task generalization capability and scene adaptability of the overall framework are effectively enhanced, and the unified and integrated modeling of cultural image understanding, semantic matching, and visual generation tasks is realized.

3. EXPERIMENTS AND RESULTS ANALYSIS

3.1 Experimental setup

In this study, the first university cultural multimodal triple dataset was constructed for model training and primary experimental validation. A total of 8,264 campus cultural images were included in the dataset, encompassing four core scene categories: historical architecture, cultural landscapes, cultural relic collections, and campus activities. The dataset was equipped with 24,792 fine-grained multi-granularity textual annotations and 15,687 structured knowledge graph triples, comprehensively covering campus cultural entities and their semantic associations. The dataset was randomly

partitioned into training, validation, and test sets at a fixed ratio of 7:1:2, with the partitioning process rigorously ensuring balanced sample distributions across scene categories and image quality levels, thereby preventing data bias from interfering with model training.

To validate the cross-domain generalization capability of the model, the publicly available cultural heritage dataset was additionally selected for generalization testing. This dataset contains over 6,000 cultural image samples, including ancient architecture, traditional cultural relics, and historical scenes, and is well-suited for the testing requirements of cross-modal modeling and image generation tasks in vertical cultural scenarios.

All experiments in this study were conducted on the Ubuntu operating system, with the hardware configuration equipped with a single NVIDIA RTX 3090 graphics processing unit. Model training was performed using the AdamW optimizer, with the base learning rate set to 2e-4, and a cosine annealing learning rate scheduling strategy was employed to achieve dynamic learning rate decay. The initial loss weights were set as $\mu_1 = 0.6$, $\mu_2 = 0.4$, $\alpha = 0.5$, $\beta = 0.3$, and $\gamma = 0.2$, and the temperature coefficient of the diffusion model was set to 0.07. The random seed was uniformly set to 42 for all experiments, and the total number of training iterations was set to 120 epochs. The model achieving the optimal performance on the validation set was retained as the final test model, thereby ensuring the reproducibility of the experimental results.

3.2 Cross-modal cultural retrieval performance validation

This experiment is designed to validate the effectiveness of the knowledge graph-guided hierarchical contrastive alignment mechanism in improving cross-modal semantic matching accuracy. Three types of retrieval tasks—image-to-text retrieval, text-to-image retrieval, and image-to-cultural-entity retrieval—were conducted on both the self-constructed university cultural dataset and the public cultural heritage dataset. The quantitative performance results of each model are presented in Figure 5.

From the experimental results, the proposed model is observed to significantly outperform the mainstream baseline models across all retrieval metrics on both datasets. On the self-constructed dataset, the Recall@1 metric of the proposed model is improved by 9.26 percentage points compared to the optimal baseline model, while Recall@5 and Recall@10 are improved by 6.63 and 6.66 percentage points, respectively, demonstrating that the hierarchical contrastive mechanism can effectively enhance the fine-grained image-text matching accuracy. On the public generalization dataset, the model maintains stable performance advantages, achieving an average improvement of over 8 percentage points compared to the baseline models, thereby validating the adaptability of structured knowledge constraints to general cultural scenes.

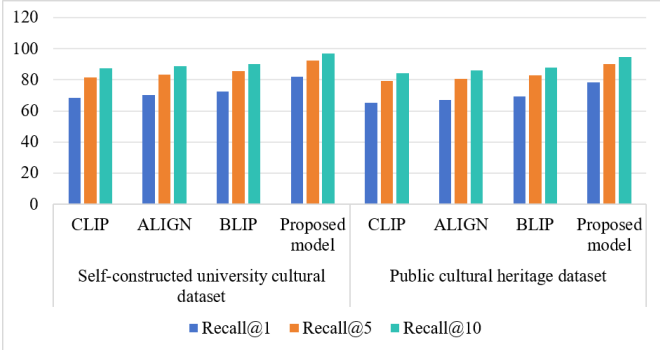

Figure 5. Cross-modal retrieval performance comparison of different models (%)

Conventional pre-trained models rely solely on sample-level global contrast for alignment, and are incapable of distinguishing between similar cultural scenes and same-category cultural entities, resulting in substantial fine-grained retrieval errors. Through the entity-level and relationship-level contrastive constraints constructed in this study, semantic matching is anchored to structured cultural entities and topological relationships, effectively resolving the matching confusion problem among similar cultural samples. This performance gain is particularly pronounced in retrieval scenarios involving small-sample niche cultural categories, demonstrating stronger fine-grained semantic retrieval capability and cross-scene generalization ability.

3.3 Cultural image enhancement effectiveness and semantic preservation validation

This experiment is designed to evaluate the visual restoration quality and cultural semantic fidelity of different models on historical campus photograph restoration and low-resolution image super-resolution tasks. Performance evaluation was conducted through a combination of visual quantitative metrics and culture-specific indicators. The experimental results are presented in Table 1.

The quantitative results demonstrate that the proposed model achieves comprehensive superiority in both visual quality and semantic preservation dimensions. Compared to the optimal baseline model, the proposed model achieves a peak signal-to-noise ratio improvement of 3.08, a structural similarity index measure improvement of 0.063, and a learned perceptual image patch similarity reduction of 0.062, indicating significantly enhanced visual texture restoration and detail reconstruction capability. On the core cultural semantic fidelity metric, the proposed model attains an accuracy of 94.28%, representing a 12.61 percentage point improvement over the baseline model, with the magnitude of performance gain substantially exceeding that observed in general visual metrics.

Table 1. Performance comparison of different models on cultural image enhancement

Model	Peak Signal-to-Noise Ratio	Structural Similarity Index Measure	Learned Perceptual Image Patch Similarity	Cultural Semantic Accuracy (%)
SRDiff	24.36	0.812	0.187	76.35
DDPM	25.18	0.835	0.162	78.92
Stable Diffusion	26.74	0.861	0.135	81.67
Proposed model	29.82	0.924	0.073	94.28

Note: SRDiff = Super-Resolution via Diffusion Models; DDPM = Denoising Diffusion Probabilistic Model.

Baseline diffusion models are optimized solely with pixel-level errors as the optimization objective, and are prone to issues such as cultural component deformation, historical texture loss, and cultural symbol alteration during the restoration process, failing to preserve semantic integrity. Through the introduced cultural semantic preservation loss, the proposed model constrains the consistency of high-level semantic features before and after restoration via a frozen culture-specific encoder. While optimizing image clarity and textural details, the cultural semantic attributes of the image are firmly maintained. Furthermore, the model exhibits stable adaptability to historical images with varying degrees of degradation. For severely blurred and old scanned low-quality images with significant distortion, the dual effects of visual optimization and semantic fidelity are still achieved, effectively addressing the application deficiencies of conventional image enhancement models in cultural vertical scenarios.

3.4 Controllable cultural image generation quality assessment

This experiment was conducted with image generation driven by dual conditions—textual descriptions and cultural knowledge entities—and model performance was quantified from three dimensions: generation quality, diversity, and cultural authenticity. The quantitative metrics and human evaluation results are presented in Table 2.

Table 2. Performance comparison of different models on cultural image generation

Model	Fréchet Inception Distance (FID)	Inception Score	Cultural Alignment	Semantic Accuracy	Visual Authenticity
Stable Diffusion	23.57	7.82	6.35	6.12	7.03
ControlNet	19.84	8.15	7.18	6.95	7.68
Diffusion Transformer	17.26	8.43	7.52	7.36	8.12
Proposed model	11.39	9.27	9.14	9.08	9.35

From the experimental data, the proposed model achieves a significantly reduced Fréchet Inception Distance (FID) score and a markedly improved inception score, demonstrating that the pixel distribution of the generated images more closely matches real campus cultural scenes, with superior image clarity and content diversity. In the human-evaluated cultural dimensions, the proposed model surpasses all baseline models by a substantial margin across all three metrics, with both cultural alignment and semantic accuracy exceeding 9.0 points, thereby achieving a high degree of matching between cultural semantics and generated content.

Generic generation models rely solely on textual semantic guidance for generation, lacking structured cultural knowledge constraints, and are highly susceptible to semantic drift issues such as cultural element disorganization, anachronistic stylistic mismatches, and entity feature deviations. Through the construction of a unified generation condition by integrating knowledge graph entity embeddings with attribute text features, and by constraining the generation process with a cultural style consistency loss, the proposed model enables precise response to fine-grained cultural semantic instructions and accurately reproduces campus architectural styles, cultural symbols, and scene layouts from different historical periods. Visual examples further corroborate that images generated by baseline models frequently exhibit issues such as missing arched doors and windows, distorted university motto symbols, and disorganized historical architectural structures, whereas the proposed model completely preserves the distinctive campus cultural features, with generation results demonstrating exceptionally high cultural authenticity and semantic controllability.

3.5 Effectiveness validation of the feature decoupling mechanism

This experiment is designed to validate the separation capability of the orthogonal decoupling network for visual morphological and cultural semantic features through quantitative cultural classification experiments, feature swapping experiments, and attention visualization experiments. The comparison of cultural classification accuracy before and after feature decoupling is presented in Figure 6.

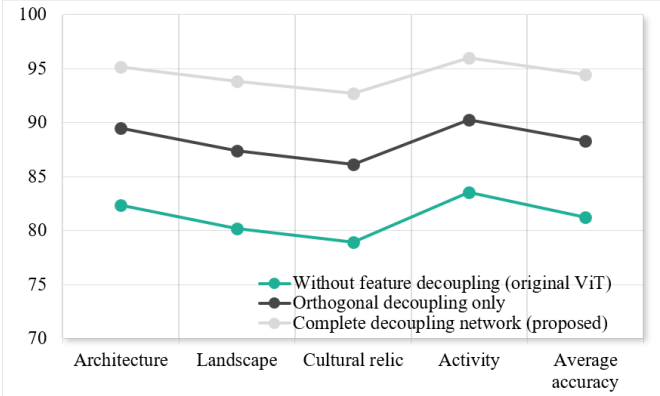

Figure 6. Comparison of cultural classification accuracy before and after feature decoupling (%)

Quantitative classification results demonstrate that the cultural feature decoupling mechanism substantially enhances the discriminative capability of semantic features. Due to feature entanglement and background noise interference, the original Vision Transformer model achieves an average cultural category classification accuracy of only 81.25%. After the introduction of orthogonal decoupling constraints alone, the model accuracy is improved by 7.07 percentage points, demonstrating that orthogonal subspace separation can effectively eliminate redundant visual noise. With the complete network incorporating the adaptive gating fusion unit, the average accuracy is further enhanced to 94.43%, with significant optimization achieved across all category-specific cultural recognition accuracies.

Feature swapping experiment results indicate that after swapping the morphological and cultural features of different images, the generated images can fully inherit the corresponding cultural semantic attributes, and the visual

morphology and humanistic semantics can be independently modulated, intuitively confirming that the two feature types are effectively decoupled. Attention visualization results demonstrate that the attention regions of the proposed culturally salient-guided encoding network are highly concentrated on core cultural areas such as inscriptions, university emblems, and landmark buildings, whereas the attention of the standard Vision Transformer is uniformly dispersed across the entire image background, lacking targeted focusing capability. The adaptive gating unit can dynamically adjust feature weights according to the semantic density of the image content: for historical cultural images, the expression of semantic features is strengthened, while for modern scene images, the preservation of visual morphological details is prioritized. These results comprehensively validate the adaptive representation advantages of the decoupling and fusion mechanisms.

3.6 Downstream task generalization performance testing

This experiment was conducted on two downstream tasks—visual question answering and image captioning—based on the unified cross-modal semantic space, with cross-dataset generalization validation also performed. The results are presented in Table 3. Accuracy was adopted as the evaluation metric for visual question answering, while Bilingual Evaluation Understudy-4 (BLEU-4) and Consensus-based Image Description Evaluation with Damping (CIDEr-D) were employed as the core evaluation metrics for image captioning.

Table 3. Downstream task generalization performance comparison

Model	Visual Question Answering Accuracy (%)	Image Captioning BLEU-4	Image Captioning CIDEr-D	Cross-Dataset Transfer Accuracy (%)
CLIP	75.32	0.612	0.725	70.15
BLIP	79.68	0.657	0.783	73.82
Proposed model	88.94	0.743	0.896	82.57

Note: BLEU-4 = Bilingual Evaluation Understudy-4; CIDEr-D = Consensus-based Image Description Evaluation with Damping; CLIP = Contrastive Language–Image Pre-training; BLIP = Bootstrapping Language–Image Pre-training.

Table 4. Core module ablation experiment results

Experimental Configuration	Retrieval Recall@1 (%)	Enhancement Peak Signal-to-Noise Ratio	Generation Fréchet Inception Distance (FID)
Complete model	81.94	29.82	11.39
Removal of the feature decoupling module	73.26	27.15	15.84
Removal of the entity-level contrastive loss	76.53	28.03	13.26
Removal of the relationship-level contrastive loss	78.12	28.67	12.51
Removal of the cultural semantic preservation loss	79.35	26.48	14.73

The experimental results demonstrate that the constructed unified semantic space possesses excellent multi-task adaptability and zero-shot generalization capability. On campus cultural downstream tasks, the proposed model achieves a visual question answering accuracy improvement of 9.26 percentage points over the optimal baseline, with substantial improvements achieved across both core image captioning metrics, demonstrating that the unified semantic space can precisely support various cultural semantic reasoning tasks. In cross-dataset transfer testing, only a minor performance degradation is observed in the proposed model, which still significantly outperforms generic cross-modal models.

The semantic spaces of generic models are not optimized for cultural scenarios, exhibiting strong modal heterogeneity, poor downstream task adaptability, and susceptibility to semantic adaptation deviations during cross-domain transfer. Through the dual constraints of distribution alignment and semantic consistency, the proposed approach unifies the feature distributions and semantic logics across four modalities, establishing a general-purpose representation system adapted to cultural scenarios that can adapt to multiple downstream tasks without fine-tuning. The slight performance degradation observed in cross-domain transfer is primarily attributed to inherent differences in the semantic systems and visual styles across different cultural scenes, which can be further mitigated through domain-adaptive optimization in future work.

3.7 Module ablation and sensitivity analysis

To validate the independent effectiveness of each core innovative module, the feature decoupling module, entity-level contrastive loss, relationship-level contrastive loss, and cultural semantic preservation loss were progressively removed in this study, and the performance variations of the model on retrieval, enhancement, and generation tasks were tested. The ablation experiment results are presented in Table 4.

The ablation results clearly validate the positive contributions of each module. The feature decoupling module exerts the greatest influence on overall performance, with substantial degradation observed in retrieval accuracy, image enhancement, and generation performance after its removal, demonstrating that purified decoupled cultural features serve as the foundational guarantee for subsequent semantic alignment and generation tasks. The entity-level and relationship-level contrastive losses directly affect cross-modal alignment accuracy, with removal resulting in diminished fine-grained semantic matching capability and indirectly reducing the semantic precision of generation and enhancement tasks. The cultural semantic preservation loss primarily affects image generation and enhancement tasks; after its removal, the visual optimization capability of the model is partially retained, but the issue of cultural semantic distortion becomes significantly prominent, with clear degradation observed in FID and peak signal-to-noise ratio metrics. Each module performs its distinct function while being mutually coupled, collectively constituting a complete

cultural intelligent modeling system in which no single component is dispensable.

Sensitivity experiments were conducted in this study on the cultural saliency coefficient λ, the hierarchical loss weights $\alpha/\beta/\gamma$, and the temperature coefficient τ, through which the optimal parameter value ranges were determined. The performance impact results of the core parameters are presented in Table 5.

The parameter experimental results indicate that the model performance exhibits significant sensitivity to all three types of core hyperparameters. When the cultural saliency coefficient is set to 0.5, the model effectively balances the feature weights between background information and core cultural regions; a value that is too small fails to effectively focus on cultural regions, while a value that is too large excessively neglects the fundamental visual information of the image. When the hierarchical loss weights are configured with a ratio of 0.5/0.3/0.2, both global sample matching accuracy and fine-grained structured semantic constraints are effectively accommodated, adapting to the hierarchical semantic characteristics of university culture. When the temperature coefficient is set to 0.07, the optimal dispersion of feature distribution is achieved, enabling effective discrimination between similar cultural samples and maximizing the semantic discriminative capability of the model. All experiments in this study are conducted with the optimal parameter combination, thereby ensuring that model performance is maintained within the optimal range.

Table 5. Key parameter sensitivity experiment results

Parameter	Value	Retrieval Recall@1 (%)	Generation Fréchet Inception Distance (FID)
λ	0.2	77.35	14.28
	0.5	81.94	11.39
	0.8	80.12	12.65
$\alpha/\beta/\gamma$	0.3/0.4/0.3	79.26	12.84
	0.5/0.3/0.2	81.94	11.39
	0.6/0.2/0.2	80.57	12.16
τ	0.05	80.23	12.41
	0.07	81.94	11.39
	0.1	79.68	13.05

Figure 7. Implementation effect diagram of university cultural content intelligent modeling based on cross-modal image processing

To validate whether the proposed method can simultaneously accomplish visual quality enhancement, cultural element localization, cross-modal semantic parsing, and semantic-preserving content generation on real university cultural images, further visualization experiments are conducted. As illustrated in Figure 7, the method first takes historical campus architecture images as input. During the enhancement stage, key visual regions—including the main building body, inscribed plaques, arched entrances, and red brick textures—are prominently highlighted, demonstrating that the cultural saliency attention mechanism can effectively suppress non-core background interference such as trees, sky, and external air conditioning units, thereby enabling the model to focus on image regions with cultural identification value. In the local parsing stage, visual segments such as plaques, window lattices, arched entrances, and brick wall textures are

mapped by the model into semantic units including text recognition, architectural type, structural features, and textural features, reflecting the fine-grained alignment capability between local image features and textual semantics. In the multi-stage processing results, structural edge extraction preserves the building facade contours and window column arrangements, salient region responses are concentrated on plaques, entrances, and main facades, and degradation simulation with semantic-preserving reconstruction intuitively demonstrates the structural fidelity capability of the model in restoring low-quality historical images. Furthermore, by combining image enhancement with cultural content generation results, it can be observed that the model, while improving clarity, restoring color, and reconstructing textural details, does not compromise the consistency of cultural symbols such as plaques, arched entrances, and red brick facades. These visualization results are consistently supported by the quantitative experimental results of this study. In cultural image enhancement tasks, the proposed model achieves a peak signal-to-noise ratio of 29.82, a structural similarity index measure of 0.924, a learned perceptual image patch similarity of 0.073, and a cultural semantic accuracy of 94.28%. In controllable cultural image generation tasks, the FID is reduced to 11.39, and the cultural alignment, semantic accuracy, and visual authenticity scores reach 9.14, 9.08, and 9.35, respectively. These results demonstrate that the proposed framework not only achieves superior visual reconstruction quality compared to generic models, but also stably preserves the architectural forms, textual symbols, and humanistic semantics of university cultural images through feature decoupling, hierarchical cross-modal alignment, and semantic-preserving diffusion generation mechanisms.

4. DISCUSSION

Within the complete framework, the feature decoupling, hierarchical semantic alignment, and semantically constrained generation modules form a complete closed-loop optimization chain, with each unit achieving gradient intercommunication and bidirectional empowerment through the unified cross-modal semantic space, without isolated independent operation. The feature decoupling network accomplishes the subspace separation of visual morphological information and cultural semantic information, eliminating the interference of irrelevant textures and background noise on cultural representations, providing high-purity semantic features for subsequent hierarchical contrastive learning, and reducing the confounding errors of fine-grained cross-modal matching at the source. Through the three-tier contrastive constraints constructed on the basis of the knowledge graph, discrete visual representations are mapped into a unified semantic space endowed with structured logic, and standardized cultural vectors are output as generation conditions for the diffusion model, constraining the content direction of image enhancement and generation processes from the semantic level. Concurrently, the semantic preservation loss embedded within the image generation and restoration branches enables the backpropagation of optimization gradients, continuously refining the front-end feature extraction and semantic alignment processes. The semantic fidelity effects of the generation tasks are used to verify the optimization degree of the representation and alignment stages, thereby forming a complete technical chain in which feature purification, semantic anchoring, and visual regeneration mutually support one another.

The performance of the proposed method is constrained by the completeness of annotation resources and the visualizability of cultural content, with clearly defined application boundaries. For niche campus cultural categories with scarce sample quantities and insufficient annotation materials, the model cannot adequately learn the representation regularities of culture-specific entities, resulting in significant performance degradation in retrieval and generation tasks. For abstract spiritual and cultural connotations, the model cannot achieve complete modeling based solely on image visual information, and still relies on accompanying textual descriptions and knowledge graph triples to provide semantic supplementation. In the absence of structured annotations, semantic alignment accuracy declines markedly. The framework possesses fundamental transferability to other cultural heritage scenarios. For vertical cultural domains such as ancient architecture and museum collections, adaptation can be accomplished by simply replacing the domain-specific knowledge graph and performing minor fine-tuning of the visual encoding layer to match the visual characteristics of the target scenario. Currently, the model covers only three modalities—image, text, and knowledge graph—and has not yet integrated multi-source cultural data such as audio and three-dimensional point clouds, leaving room for expansion in multimodal coverage.

From the perspective of digital humanities and the practical implementation of university cultural digitization projects, the integrated cross-modal modeling system constructed in this study fills the specialized technical gap in intelligent cultural image processing and possesses multi-level engineering application value. Through the capabilities of the framework, tasks such as batch restoration of historical campus photographic archives, intelligent retrieval of campus cultural resources, and automatic generation of customized campus cultural scenes can be accomplished, supporting the development of digital products including digital university history museums, online cultural exhibition halls, and general education cultural courseware. Traditional university cultural digitization efforts have largely remained at the level of material storage and simple image beautification, without achieving deep semantic parsing or controllable regeneration of cultural content. The proposed method bridges the integration pathway between visual image processing and structured cultural knowledge, completely preserving historical symbols, architectural forms, and humanistic connotations while ensuring visual output quality, providing a standardized intelligent solution for the long-term digital preservation and lightweight online dissemination of university cultural resources, while also offering a reusable technical paradigm for vertical-scenario modeling at the intersection of digital humanities and computer vision.

Furthermore, the proposed framework provides an extensible pathway for the institution-specific development of distinctive university cultural resources. For application-oriented higher education institutions with distinctive regional cultural characteristics, industrial backgrounds, or educational traditions, visual resources such as local historical architecture, industrial memory, regional cultural landscapes, and university-industry collaboration scenes can be incorporated into the unified cross-modal semantic space. Through image enhancement, semantic annotation, knowledge association, and content generation, scattered local

cultural materials can be transformed into teaching cases, digital exhibition resources, virtual simulation materials, and campus cultural dissemination content. This process enables bidirectional transformation among local resources, university culture, and digital teaching resources, enhancing the scene adaptability, institutional distinctiveness, and application promotion value of university cultural content modeling.

5. CONCLUSION AND FUTURE DIRECTIONS

To address the critical challenges of intelligent image modeling in university cultural digitization projects, a cross-modal intelligent modeling framework for cultural semantic preservation was constructed in this study, through which three core technical problems were systematically resolved: the entanglement of cultural visual representations, insufficient precision in fine-grained cross-modal semantic alignment, and semantic drift during image generation and enhancement stages. Within the framework, low-level visual morphology and high-level cultural semantics were separated through an orthogonally constrained feature decoupling network; a three-tier progressive contrastive learning system encompassing samples, entities, and relationships was constructed on the basis of a university cultural knowledge graph, through which structured semantic alignment was achieved; and standardized cultural semantic vectors were embedded as conditions into the diffusion model with an additional dedicated semantic preservation loss, forming a closed-loop technical chain from feature purification and semantic mapping to visual regeneration. Through multiple sets of comparative, ablation, and generalization validation experiments conducted on the self-constructed university cultural multimodal triple dataset and the public cultural heritage dataset, both quantitative metrics and qualitative visualization results collectively demonstrated that the proposed method significantly outperformed existing state-of-the-art models in cross-modal retrieval, historical image restoration, and controllable cultural scene generation tasks, simultaneously achieving both visual quality optimization and complete preservation of cultural semantics. This study advances the theoretical implementation pathway of cross-modal visual modeling in vertical digital humanities scenarios, provides a complete technical solution for the digital preservation and intelligent dissemination of university campus culture, and offers a reference-worthy technical paradigm for cultural heritage image processing research.

Given the application limitations of the current framework, future work can be extended along three dimensions. First, multicultural media such as audio, short-form video, and three-dimensional point clouds can be incorporated to expand the modality adaptation scope of the unified cross-modal semantic space, enabling integrated modeling of omnimedia campus cultural resources. Second, few-shot and zero-shot semantic representation learning schemes can be explored to reduce the dependence of the model on large-scale manual annotations and to lower the data construction costs of cultural digitization projects. Third, the two-dimensional image generation capability can be extended to three-dimensional campus scene reconstruction and interactive virtual tour content generation, enriching the presentation forms of digital cultural resources and further broadening the practical application scenarios of this modeling framework in the digital humanities domain.

ACKNOWLEDGEMENT

This paper was supported by Construction and Practice of a New Digital-Intelligent Education and Teaching Ecosystem in Application-Oriented Undergraduate Universities from the Perspective of Quality Culture (Grant No.: 25BG084) and Establishment of an Ecosystem for Deep Integration Between Artificial Intelligence and Education Based on the Coordination of Technology, Educators and Learners (Grant No.: 25SJK02).

REFERENCES

[1] Lian, Y., Xie, J. (2024). The evolution of digital cultural heritage research: Identifying key trends, hotspots, and challenges through bibliometric analysis. Sustainability, 16(16): 7125. https://doi.org/10.3390/su16167125

[2] Chunlan, Y., Tengku Wook, T.S.M., Rosdi, F. (2025). Advancing cultural heritage: A decadal review of digital transformation in Chinese museums. NPJ Heritage Science, 13: 189. https://doi.org/10.1038/s40494-025-01714-x

[3] Buragohain, D., Meng, Y., Deng, C., Li, Q., Chaudhary, S. (2024). Digitalizing cultural heritage through metaverse applications: Challenges, opportunities, and strategies. Heritage Science, 12: 295. https://doi.org/10.1186/s40494-024-01403-1

[4] Hu, J., Yu, Y., Zhou, Q. (2025). GuidePaint: Lossless image-guided diffusion model for ancient mural image restoration. NPJ Heritage Science, 13: 118. https://doi.org/10.1038/s40494-025-01693-z

[5] Li, Y., Zhang, C., Li, Y., Sui, D., Guo, M. (2025). An improved mural image restoration method based on diffusion model. NPJ Heritage Science, 13: 347. https://doi.org/10.1038/s40494-025-01914-5

[6] Zou, J., Du, Y., Liu, G., Jiao, Z., Zhang, H. (2025). Generating Chinese intangible cultural heritage images with structure and color awareness. NPJ Heritage Science, 13: 579. https://doi.org/10.1038/s40494-025-02150-7

[7] Zhang, J., Huang, J., Jin, S., Lu, S. (2024). Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(8): 5625-5644. https://doi.org/10.1109/TPAMI.2024.3369699

[8] Chen, F.L., Zhang, D.Z., Han, M.L., et al. (2023). VLP: A survey on vision-language pre-training. Machine Intelligence Research, 20(1): 38-56. https://doi.org/10.1007/s11633-022-1369-5

[9] Lymperaiou, M., Stamou, G. (2024). A survey on knowledge-enhanced multimodal learning. Artificial Intelligence Review, 57(10): 284. https://doi.org/10.1007/s10462-024-10825-z

[10] Zhu, X., Li, Z., Wang, X., et al. (2024). Multi-modal knowledge graph construction and application: A survey. IEEE Transactions on Knowledge and Data Engineering, 36(2): 715-735. https://doi.org/10.1109/TKDE.2022.3224228

[11] Liang, W., De Meo, P., Tang, Y., Zhu, J. (2024). A survey of multi-modal knowledge graphs: Technologies and trends. ACM Computing Surveys, 56(11): 273. https://doi.org/10.1145/3656579

[12] Hogan, A., Blomqvist, E., Cochez, M., et al. (2021).

Knowledge graphs. Association for Computing Machinery Computing Surveys, 54(4): 1-73. https://doi.org/10.1145/3447772

[13] Xu, Z., Zhang, X., Chen, W., et al. (2023). A review of image inpainting methods based on deep learning. Applied Sciences, 13(20): 11189. https://doi.org/10.3390/app132011189

[14] Jiang, R., Zheng, G.C., Li, T., Yang, T.R., Wang, J.D., Li, X. (2024). A survey of multimodal controllable diffusion models. Journal of Computer Science and Technology, 39(3): 509-541. https://doi.org/10.1007/s11390-024-3814-0

[15] Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M. (2023). Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9): 10850-10869. https://doi.org/10.1109/TPAMI.2023.3261988

[16] Cao, H., Tan, C., Gao, Z., et al. (2024). A survey on generative diffusion models. IEEE Transactions on Knowledge and Data Engineering, 36(7): 2814-2830. https://doi.org/10.1109/TKDE.2024.3361474

[17] Zhao, F., Ren, H., Su, Z., Zhu, X., Zhang, C. (2025). Diffusion-based heterogeneous network for ancient mural restoration. NPJ Heritage Science, 13: 206. https://doi.org/10.1038/s40494-025-01719-6

[18] Wu, J., Tian, H., Yan, W. (2025). Application of GANs in ancient architectural heritage image restoration. NPJ Heritage Science, 13: 655. https://doi.org/10.1038/s40494-025-02234-4