npj | heritage science

Article

https://doi.org/10.1038/s40494-025-01981-8

Virtual restoration method of Kizil Grotto murals based on multimodal controlled diffusion models

Check for updates

Guoyan Lv[1], Huiqin Wang[1] ✉, Ke Wang[1], Huaidong Zhao[2] & Li Zhao[3]

Generative deep learning provides new approaches for natural image restoration and structural reconstruction. However, virtually restoring Kizil cave murals remains difficult due to their unique artistic style, complex damage, and the need to preserve semantic consistency. This study proposes a multimodal controlled diffusion model that integrates textual and multi-dimensional visual features for high-precision restoration. The model leverages latent space diffusion for high-quality image generation and introduces structural constraints to improve semantic alignment and controllability. A dynamic feature-adaptive GSC (DFA-GSC) module captures local and global features through multi-scale convolution and an adaptive weight generator, enhancing texture perception. For damaged regions, a conditional matching loss helps refine both texture and structure. Experimental results demonstrate that compared to traditional CNNs and single diffusion models, the proposed method achieves superior performance in both evaluation metrics and visual quality.

The murals of the Kizil Grottoes are not only invaluable treasures of Buddhist art but also crucial evidence for the study of history, religion, and cultural exchange[1]. The murals of the Kizil Caves, with their distinctive cave forms and painting styles, reflect the eastward transmission of Buddhism through the Western Regions and reveal the cultural interactions between East and West as well as the trajectory of Sinicization. However, natural erosion, climatic fluctuations, and human activities—such as large-scale looting by foreign expeditions, shifts in local religious beliefs, and the scraping of gold foil from mural surfaces—have led to varying degrees of degradation, particularly in facial regions, posing serious challenges to cultural heritage preservation.

In early mural restoration practices, the quality of restoration heavily relied on the subjective judgment and artistic skill of individual restorers, making it difficult to ensure scientific rigour and consistency. Moreover, the traditional cultural heritage restoration sector faces issues such as an imbalance between supply and demand, a shortage of skilled professionals, and a high technical entry barrier. Consequently, museums worldwide are increasingly adopting digital image restoration technologies for virtual restoration, which has become a key area of research in cultural heritage conservation[2,3].

Traditional image restoration techniques mainly include diffusion-based restoration methods and sample block matching-based restoration methods. Diffusion-based techniques restore missing areas by propagating pixel information from surrounding regions into the damaged zones using predefined diffusion functions[4–6]. Chen Yong et al.[7] proposed an improved curvature-driven diffusion algorithm tailored to the crack restoration of Dunhuang murals, which optimises the diffusion term in the transitional regions around crack edges, enhancing restoration fidelity.

Exemplar-based approaches leverage image redundancy by selecting similar patches within the image to fill in the missing regions[8–14]. Representative works include the block-matching method introduced by Criminisi et al.[8], the random sampling-based PatchMatch algorithm by Barnes et al.[11], and the sparse representation-based enhancement strategies developed by Shen et al.[12], all of which have been widely adopted in the field of cultural heritage restoration[13,14]. The main idea of these methods is to complete the repair task by iteratively performing three steps: similarity calculation of the sample blocks to be repaired, searching for the best matching sample blocks and filling, which is more suitable for repairing local damage areas such as small cracks, scratches and scribbles. However, due to the lack of global context modelling capability, it is inadequate for large-scale mural restoration tasks with complex structures and high requirements for stylistic consistency.

The advent of deep learning provides new possibilities for image restoration. Currently, deep learning-based restoration approaches can be broadly categorised into non-guided and guided methods. Non-guided restoration primarily relies on the inherent features of the damaged image for reconstruction. For instance, Zeng et al.[15] proposed a pyramid context encoder that restores image content via hierarchical transfer of deep and

[1]College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an, China. [2]College of Arts, Xi'an University of Architecture and Technology, Xi'an, China. [3]Kizil Caves Research Institute, Xinjiang Uygur Autonomous Region, Kizil, China. ✉e-mail: hqwang@xauat.edu.cn

shallow features. Li et al.[16] introduced a visual structure reconstruction network that integrates reconstruction layers into the encoder–decoder framework to enhance structural fidelity. Suvorov et al.[17] expanded the receptive field using fast Fourier convolution to capture periodic structural patterns, but lacked the boundary semantic constraints. Zheng et al.[18] developed a two-stage restoration pipeline that first generates coarse results, followed by a refinement stage, offering improved performance in localised reconstruction. While effective to some extent, these methods are limited by their exclusive reliance on pixel-level information, often resulting in inconsistent semantics in the restored regions.

In contrast, guided restoration methods incorporate prior knowledge —such as structural edges, semantic cues, or external reference images—to inform the inpainting process. Nazeri et al.[19] proposed the EdgeConnect framework, which first reconstructs edge maps to guide content restoration. Zhao et al.[20] introduced the concept of cross-image guided restoration, and proposed the use of image chunks to guide the restoration of target images. Zhang et al.[21] utilised text semantics in their TDANet model to enhance semantic fidelity, though with limited capacity to recover fine details. Guo et al.[22] addressed distortions arising from inadequate interaction between texture and structure, proposing a texture–structure coupling network that employs a bidirectionally gated feature fusion module and a contextual aggregation mechanism to optimise detail consistency. Wan et al.[23] proposed a two-stage CNN-transformer hybrid model that first reconstructs global structure via a Transformer and then refines textures through convolutional upsampling.

Compared to natural images, cultural heritage mural images exhibit unique characteristics such as complex scenes, blurred textures, and diverse artistic styles. Deep learning-based mural restoration methods have shown promising results. For example, Xu and Fu[24] effectively restored the colours of ancient murals using a DenseNet-based algorithm. Peng et al.[25] proposed a content-constrained convolutional neural network with multi-scale feature extraction to restore murals, although this method mainly targets small-scale breakage. Yang et al.[26] introduced the 3M-Hybrid restoration model to address challenges such as data scarcity and large-scale breaks. Wang et al.[27] developed a novel dual-aggregated GAN architecture, combining multi-scale feature fusion with a polarised self-attention mechanism, achieving high-quality restoration of high-resolution mural images. However, most of these methods do not adequately account for spatial structure, making it difficult to produce satisfactory results when restoring mural images with large-scale breaks in facial regions.

Recently, diffusion models have gained increasing attention in the field of image generation, and some studies have explored their application in image restoration. Wang et al.[28] conducted a comprehensive review of diffusion models in the field of image editing, providing a theoretical foundation for understanding their potential advantages in image restoration tasks. Lugmayr et al.[29] used a pre-trained unconditional denoising diffusion probabilistic model (DDPM) as a generative prior, enabling conditional sampling without modifying the original network architecture. Meng et al.[30] proposed SDEdit, which significantly enhanced restoration quality by optimising the sampling process. Wang et al.[31] developed DDNM, refining the zero-space content during the inversion process to reduce semantic errors. Xia et al.[32] introduced DiffIR, which integrates an IR prior extraction network with a dynamic transformer, achieving efficient restoration through a two-stage training strategy. Zhang et al.[33] proposed M2S, employing a coarse-to-fine sampling strategy to greatly improve restoration efficiency. Saharia et al.[34] developed the Palette method, which performs exceptionally well across multiple restoration tasks, including image repair and super-resolution. Huang et al.[35] introduced wavelet transforms into diffusion models, constructing multi-scale frequency representations to enhance local sensitivity and improve the model's performance in detail reconstruction and texture preservation. Furthermore, Huang et al.[36] proposed a dual-schedule inversion mechanism that enables high-quality image editing without the need for additional training or fine-tuning.

Despite these advancements, the direct application of existing models —primarily trained on natural images—to mural restoration remains challenging. This is largely due to substantial differences in data distributions, structural complexity, and semantic content between natural images and cultural heritage artefacts such as the murals of the Kizil Caves. These challenges can be summarised as follows:

(1) Restoration must address pixel-level detail while preserving the murals' distinctive artistic style, colour hierarchy, and painting techniques to maintain historical authenticity.
(2) As illustrated in Fig. 1, damage in murals often follows irregular patterns, unlike the structured gaps in natural images, making it difficult to infer missing content from surrounding regions.
(3) Conventional U-Net-based architectures use fixed convolutional kernels, limiting their adaptability to the diverse features of mural images and restricting the joint modelling of multi-scale information.
(4) Deep learning approaches require large, high-quality datasets, yet there is currently no publicly available dataset of Kizil cave murals, nor paired 'damaged-complete' samples, thereby increasing the difficulty of virtual restoration.

To address these limitations, this study proposes a mural restoration framework based on a multimodally controlled diffusion model, tailored specifically for the restoration of highly stylised and intricately damaged Kizil Cave murals. The key contributions of this work are as follows:

(1) Proposing a FEDR restoration network to match suitable reference images with structural and texture similarity, and to enhance the consistency and style adaptation ability of local restoration by a double bootstrapping mechanism. The DFA-GSC module is employed to capture both global and local mural features using adaptive multi-scale convolution, thereby improving the modelling of complex textures.
(2) Designing a feature extraction algorithm that integrates structural, textural and frequency information to make up for the inadequacy of a single edge detector for detail extraction, and improve the ability of acquiring structural and textural information of the face region.
(3) A dedicated dataset for Kizil Cave mural restoration is constructed, addressing the lack of publicly available training resources.
(4) Introducing conditional matching loss to mitigate the potential error caused by the mismatch between the generated results and the control conditions, and optimising the training process.

Methods

This section focuses on the training and testing datasets used in this study, as well as the research approach and methods.

Dataset construction

To support research on the digital restoration of the Kizil Grottoes murals, this paper constructs a multi-source mural image dataset that includes original mural images, segmented damage masks, and multi-dimensional feature fusion images. The goal is to enhance model performance in style representation and structural reconstruction. The dataset creation process is divided into several stages, from data collection to organisation, with each step carefully designed to ensure high quality and practical usability. A detailed overview of the steps is provided below.

The dataset comprises 10,277 original mural images, which were collected through field photography by our laboratory researchers and provided in part by the Kizil Grottoes Research Institute of the Xinjiang Uygur Autonomous Region. All images underwent data cleaning, which involved identifying and removing duplicate or highly similar images, resizing them, and cropping irrelevant regions to focus on the core content of the Kizil murals. In the integrity screening phase, we evaluated three key aspects: colour preservation, shape contour integrity, and clarity of semantic content. We selected mural areas that were relatively complete both visually and semantically to improve the overall quality and utility of the dataset, as illustrated in Fig. 2, which depicts the data construction pipeline of the Kizil mural dataset.

This study also defines "breaks" in the murals with clear criteria, including human-caused surface scraping, large-scale pigment peeling,

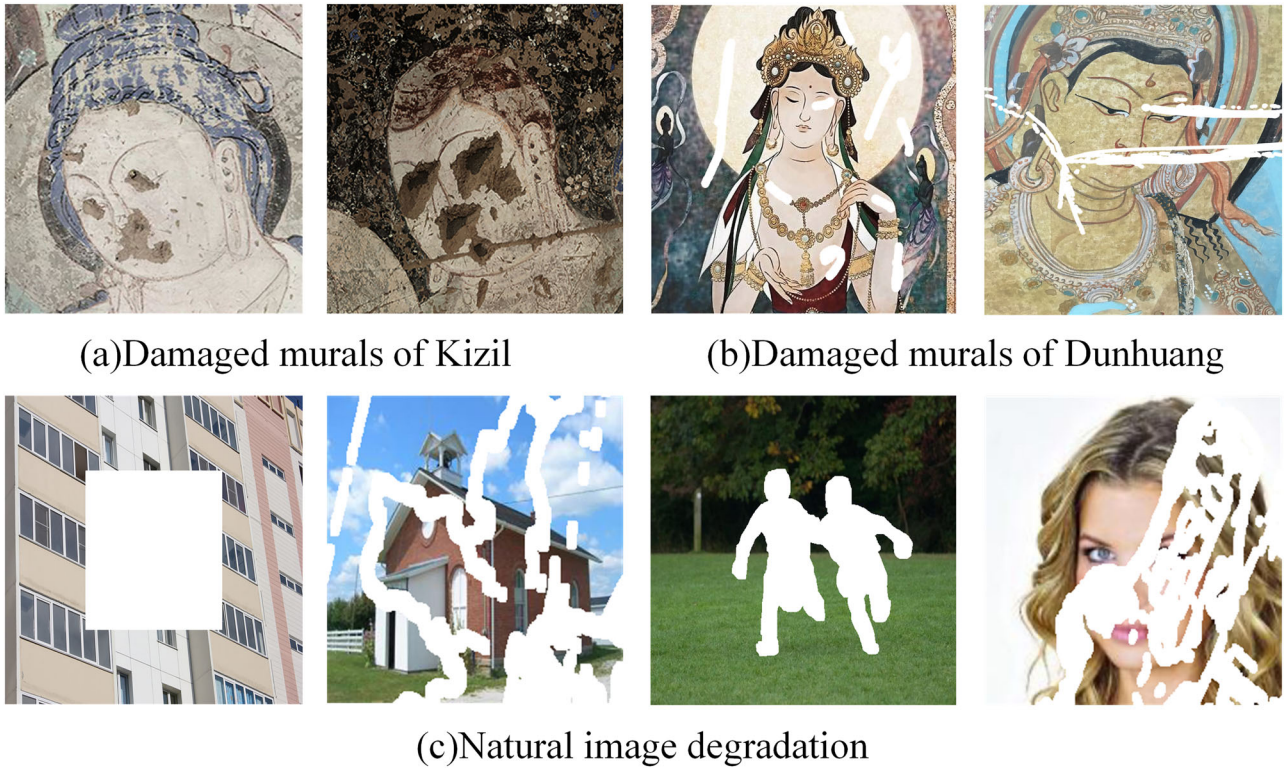

Fig. 1 | Damage patterns in mural and natural images. a Kizil Cave murals, showing complex textures and irregular, continuous damage. **b** Dunhuang murals, exhibiting similar complex and irregular damage patterns. **c** Natural images, with more regular and often discontinuous damaged regions.

cracks, discolouration, and fractures consistent with the natural ageing process of cultural heritage artefacts. This definition is based on a comprehensive analysis of the diverse and complex break patterns found in the Kizil murals. For areas with severe breaks and significant loss of structural information, we manually annotated representative missing regions with high precision. These annotations are shown in the mask dataset in Fig. 2. Given the irregularity and complexity of mural breakage, the annotation process included multiple rounds of iterative review to identify and correct omissions or errors. Each review cycle aimed to improve the accuracy and completeness of the annotations. After thorough quality control, the finalised mask data was exported in PNG format to form the segmented mural break dataset. These manually annotated break regions simulate real-world damage, allowing for the construction of paired "broken–intact" images to support supervised learning tasks in mural restoration.

Additionally, this paper generates multi-dimensional feature fusion images that correspond one-to-one with the original mural images, as shown in the multi-dimensional feature fusion dataset in Fig. 2. These images provide additional structural and stylistic guidance during image generation. Generated using the fusion algorithm detailed in Table 1, the fused features are designed to improve the model's capacity for multi-scale semantic representation and fine-grained detail reconstruction, thereby providing more precise guidance for the restoration of complex mural content.

FEDR overall network architecture

To address the significant domain gap between natural images and mural paintings in terms of colour distribution, texture patterns, and structural organisation, we propose a dual-guidance network, FEDR, tailored for the restoration of Kizil Grotto murals, as illustrated in Fig. 3. FEDR is trained on a diverse collection of mural image data and incorporates multi-modal conditions to embed both structural and semantic priors. This design helps to alleviate the adverse effects caused by cross-domain feature distribution discrepancies, thereby improving the model's adaptability and generation quality in mural-specific restoration tasks.

The architecture of the network is composed of several key components, including a text-conditional encoder[37], a multi-feature fusion module, a structural guidance branch[38], and a variational autoencoder (VAE)[39]. These components work together to enable dual-guided restoration modelling, particularly targeting the facial regions of the Kizil Grottoes murals.

The framework takes as input $I_{\text{mask}} \in \mathbb{R}^{H\times W\times 3}$ mural images combined with their corresponding binary masks, effectively integrating the original mural appearance with explicit structural and damage information. This fusion embeds spatial cues directly into the input, enabling the model to distinguish between intact and damaged regions from the very beginning. A VAE encoder is then employed to extract latent image features $\mathbf{z}_t$ from this fused representation, capturing both visual texture and structural cues necessary for accurate restoration. The mask regions explicitly highlight areas that require reconstruction, thereby sharpening the network's focus and guiding the generation process toward selectively restoring only the damaged portions of the mural. At the same time, text prompts Prompt $\in \mathbb{R}^d$ describing the mural's semantic content—such as facial features and stylistic cues—are processed by a text encoder, which converts them into semantic embedding vectors T_{emb}. These embeddings guide the generation process toward semantically coherent and contextually appropriate reconstructions.

The aforementioned mural control images and textual information are jointly fed into the main U-Net architecture, which performs guided denoising and generation through multiple Cross-Attention modules. To enhance the network's capability in detail restoration, a structurally asymmetric submodule stacking strategy is adopted within the U-Net. Both the encoder and decoder are divided into three processing stages: in each encoder stage, two CrossAttnDownBlock modules are stacked to progressively extract contextual features; in contrast, each decoder stage stacks three CrossAttnUpBlock modules to strengthen the generative modelling capacity along the upsampling path.

Inspired by the design philosophy of ControlNet[38], FEDR further incorporates a structural control path into the U-Net framework. Leveraging weight sharing, this path integrates structural prompt images I_{control}—

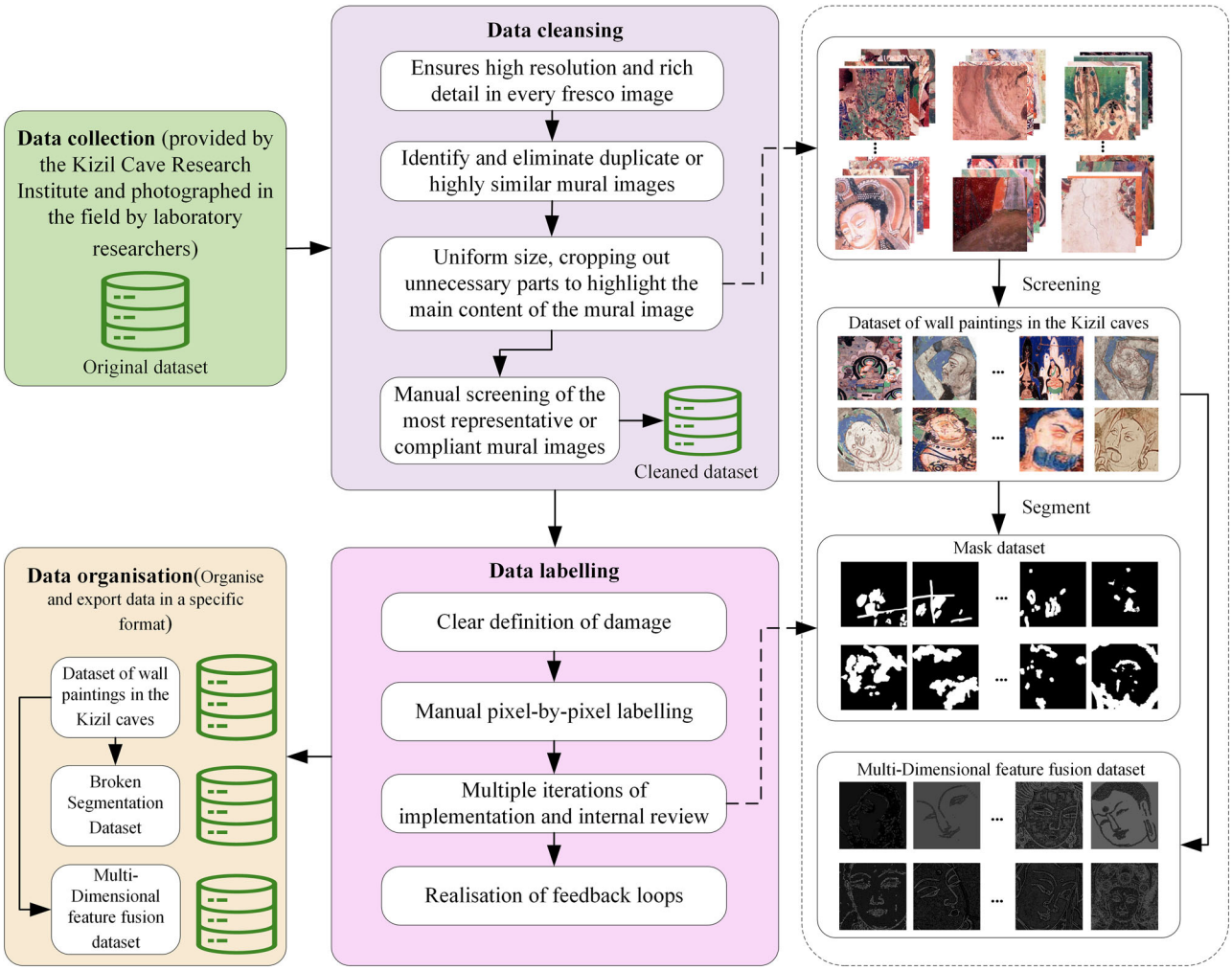

Fig. 2 | Diagram illustrating the dataset construction process.

generated by multiple feature fusion modules—to guide the restoration process in preserving key structural elements of the mural. Specifically, the structurally guided features are first spatially aligned using Zero Convolution, and then injected into the backbone network through each Cross-Attention module, thereby imposing structural constraints throughout the generation process. The computation within the Cross-Attention module is defined as follows:

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

Among these, Q, K, and V, represent the query, key, and value matrices, respectively, and d_k denotes the dimensionality of the key. To accommodate the distinct information sources for semantic and structural guidance, we designed a differentiated attention input strategy within the dual-path architecture. In the main U-Net path, the query is derived from the features of the masked mural image, while the key and value matrices are composed of semantic embeddings, enabling semantic-level guidance. In the structural control path, the query is shared with the main path, whereas the key and value matrices are generated from the control image features, providing fine-grained structural constraints. Finally, in the MidBlock layer, semantic and structural guidance signals are unified and fused, allowing the generative trajectory in latent space to be iteratively optimised—thus achieving both semantic consistency and structural coherence in mural image restoration and reconstruction.

Compared to natural images, mural images from the Kizil Grottoes exhibit greater complexity and diversity in semantic structure, textural details, and damage patterns, thereby posing more stringent demands on the feature modelling capacity of restoration networks. Although the graph structure convolution (GSC) module commonly used in diffusion models offers a concise architectural design, its fixed receptive field and static channel response mechanism limit its ability to capture multi-scale semantic structures, particularly in the context of highly degraded mural images.

To address the limited region-specific feature extraction capacity of standard U-Net architectures when applied to Kizil mural restoration, we propose an enhanced GSC variant, termed dynamic feature-adaptive improved gated spatial convolution (DFA-GSC) module[40]. This module is designed to improve the model's capability to represent large-scale contextual information and to guide the network toward more effective attention to semantically critical regions.

As illustrated in Fig. 4, the DFA-GSC module comprises two main components: a multi-scale convolutional branch and an adaptive convolutional weight generator. Input features are processed through three parallel convolutional branches to capture hierarchical spatial information:

A 1×1 convolution branch captures inter-channel dependencies;

A 3×3 convolution branch focuses on local feature extraction;

A 5×5 convolution branch captures broader contextual features via a larger receptive field.

Table 1 | The process of multi-feature fusion

Algorithm: multi-feature fusion for Kizil caves murals
Require:
Input image tensor$X$, low threshold τ_{low}, high_threshold τ_{high}, Kernel size k, Gaussian Sigmaσ, hysteresis flag h, small constant ò.
Ensure:
Fused feature map F; detected edges E.
1: Convert to grayscale
$X_{gray} = \text{rgb\_to\_grayscale(X)}$
2: Apply Gaussian Blur
3: Compute spatial gradients
4: Calculate gradient magnitude and angle
$M = \sqrt{G_x^2 + G_y^2 + \epsilon}$ $A = \text{atan2}(G_y, G_x) \times \frac{180}{\pi}$ $A = \text{round}(A/45) \times 45$
5: Non-maximum suppression (NMS)
6: Extract Gabor features
$F_{gabor} = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\cos(2\pi\frac{x'}{\lambda}+\psi)$
7: Extract HOG features
$F_{hog} = \nabla I = \left[\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}\right]$
8: Extract LBP features
$F_{lbp} = \sum_{p=0}^{P-1} s(i_p - i_c)\cdot 2^p \text{ where } s(x) = \begin{cases} 1 x \geq 0 \\ 0 x<0 \end{cases}$
9: Fuse features with weighted sum
$F = 0.1 \times F_{gabor} + 0.1 \times F_{hog} + 0.25 \times F_{lbp} + 0.55 \times M_{nms}$
10: Detect edges with thresholding
11: Return F, E.

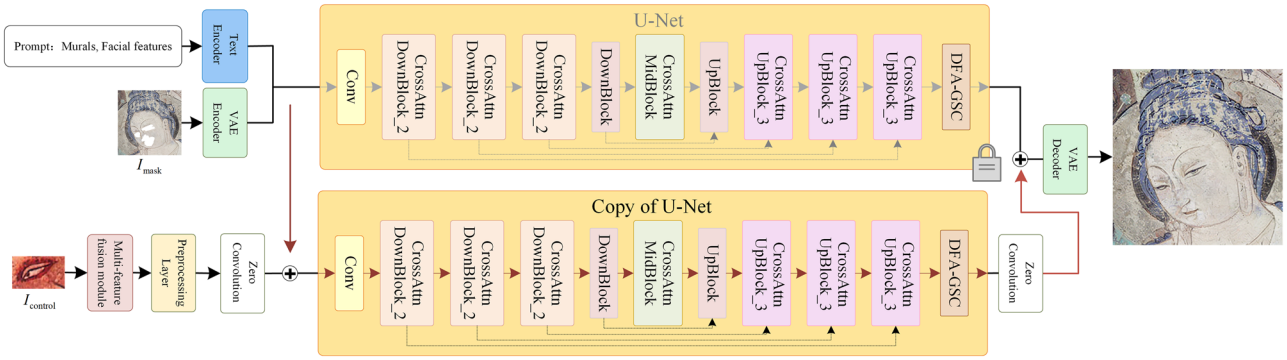

Fig. 3 | FEDR network framework for dual-guided mural restoration.

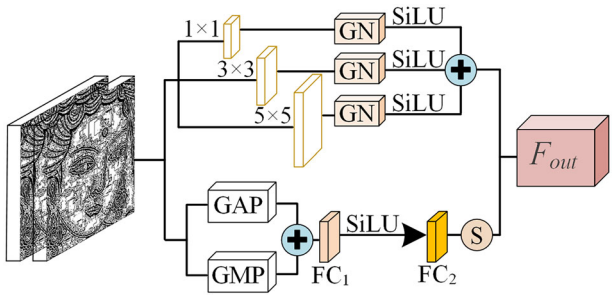

Fig. 4 | DFA-GSC module.

The outputs from the three branches are concatenated along the channel dimension to form a unified feature map $F_{ms} \in \mathbb{R}^{(C_1+C_2+C_3)\times H\times W}$. This is followed by Global Average Pooling (GAP) and Global Max Pooling(GMP)[41] to generate two global descriptors $F_{gap}, F_{gmp} \in \mathbb{R}^C$. The pooled descriptors are then passed through two shared fully connected layers to reduce the channel dimension C to C/r (where r is the reduction ratio), and then restored back to C to produce the dynamic convolutional weights:

$$W_{\text{dynamic}} = \sigma(FC_2(SiLU(FC_1(F_{gap} + F_{\text{gmp}})))) \quad (2)$$

Table 2 | Numerical characteristics of facial features in varying regional and temporal styles

Different regional styles	Gabor features	LBP features	HOG features
Early Kizil murals	0.971	1.00	1.00
Mural paintings from the Dunhuang Middle Period	0.540	0.074	0.499
Han dynasty tomb mural paintings	0.318	0.504	0.221
Tang dynasty tomb mural paintings	0.294	0.331	0.175
Yuan dynasty temple and monastery murals	0.464	0.000	0.000

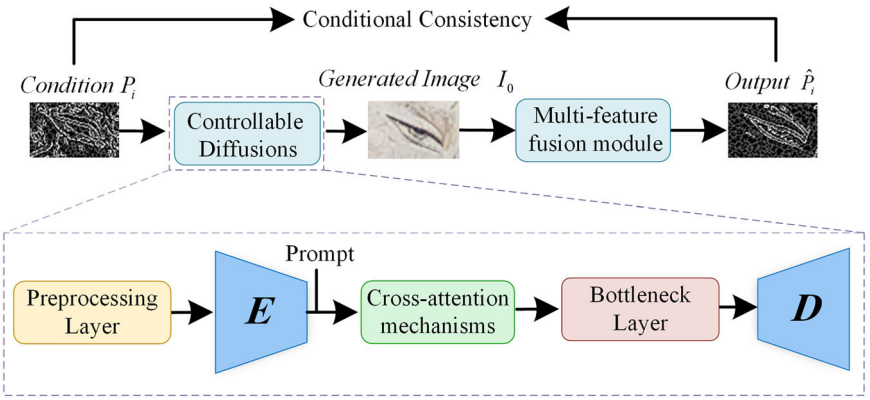

Fig. 5 | Circular consistency diagram.

Finally, these adaptive weights are applied to reweight and fuse the multi-scale feature map, yielding the enhanced output:

$$F_{\text{fused}} = W_{\text{dynamic}} \otimes F_{ms} \tag{3}$$

where $\otimes$ denotes channel-wise multiplication and σ is the sigmoid activation function.

To address the challenge of restoring missing facial regions in the murals of the Kizil Caves, this paper proposes a restoration guidance approach based on multi-dimensional feature fusion. Specifically, the multi-dimensional feature fusion module (as illustrated in Fig. 3) is employed to implement the style-guided feature extraction described in this section. The core objective is to construct a multi-dimensional fused feature map that balances stylistic distinctiveness and structural stability, providing prior style guidance for subsequent restoration processes.

First, considering that the Kizil Grottoes murals exhibit distinct regional stylistic characteristics, this study systematically extracts and normalises mural figure images from various regional styles across multiple feature dimensions. As shown in Table 2, the analysis results indicate that the Kizil murals demonstrate stylistic separability within the following feature spaces.

To enhance structural control during the restoration process, we further introduce a multi-dimensional feature fusion algorithm that integrates Gabor, HOG, LBP, and Canny descriptors through a weighted combination. This fusion provides a more comprehensive representation of facial contours and fine-grained texture details. Feature normalisation is applied to improve both the expressiveness and stability of the representation. The resulting fused map is embedded as a conditional input into the dual-guided generative network, serving as an auxiliary structural prior to guide the reconstruction of missing mural regions. This approach is particularly effective in cases of extensive damage or severe semantic degradation, where textual prompts alone are insufficient. By compensating for the lack of reliable structural cues, the fused feature prior plays a crucial role in enhancing restoration quality in complex or degraded areas.

To enhance the control capabilities and consistency of the generated results of multimodal diffusion models, we propose a recurrent consistency training strategy that enables the model to generate images that are structurally faithful and semantically reasonable under the joint guidance of text prompts and visual control conditions. This mechanism establishes a bidirectional mapping between the input multi-dimensional feature fusion conditions and the generated mural images, ensuring that the generated results are faithful to the input conditions in both structural and semantic aspects, thereby significantly improving the controllability and consistency of the generated mural images.

Specifically, first, the optimal Kizil mural reference image fusion features are given as input control conditionsP_i. Combined with text promptsP_t, mural images I_0 are generated through a controllable diffusion model. Subsequently, a multi-feature fusion model is used to re-extract multi-feature fusion conditions $\hat{P}_i$ from the generated mural imagesI_0, and these are matched with the input conditionsP_i to minimise conditional loss, thereby achieving a cyclic consistency constraint, as shown in Fig. 5. Ideally, condition $P_i \approx \hat{P}_i$ should be satisfied, meaning that the mapping process from conditions to generated images and back to conditions remains consistent, thereby ensuring that the input conditions effectively guide the generation process.

To implement the above mechanism, the pre-processing module first extracts features from the input multi-dimensional feature fusion control image through two 3×3 convolution layers and ReLU activation functions. Then, it stabilises the feature representation at the beginning of model training through zero convolution layers and maps it to a spatial feature map:

$$f_{pre} = \text{ZeroConv}(\text{ReLU}(\text{Conv}(P_i))) \tag{4}$$

Subsequently, the encoder component extracts high-level visual features through convolution. The cross-attention mechanism combines text prompt features with visual features, using an attention weight distribution mechanism to map text embeddings as Key and Value, with control diagram features as Query, thereby achieving deep integration of multi-modal semantic information. After feature compression and decoder upsampling, and deconvolution operations, the final target mural image I_0 is generated.

During training, a condition-matching loss based on multi-feature fusion was designed, combined with a time step weighting mechanism to dynamically adjust the weight distribution between the noisy early stages

Fig. 6 | Restoration results of different algorithms on simulated broken mural paintings. Columns show: (**a**) Input damaged murals, (**b**) ground truth, (**c**) Repaint [28], (**d**) CTSDG [22], (**e**) ICT [23], (**f**) SD [40], (**g**) BrushNet [45], (**h**) MuralNet [46], and (**i**) Our method.

and the clear later stages of the image, in order to avoid noise interference and enhance the generation of high-quality images. The condition matching loss[38] is defined as:

$$L_{\text{condition}} = \sum_{t=t_{\text{thre}}}^{1} \lambda_t \parallel E(G(P_t, P_i)) - P_i \parallel_2^2 \tag{5}$$

Where

$E(\cdot)$ denotes the multi-dimensional feature fusion extraction function, which is applied to both the generated mural image and the reference control image to map them into the semantic feature space for consistency comparison;

$G(\cdot)$ denotes the diffusion model;

λ_t is the weight coefficient of timestep t, dynamically adjusted to reflect its relative importance: smaller weights are assigned at early timesteps (larger t) when noise is dominant, and larger weights at later timesteps (smaller t) when the image becomes clearer;

t_{thre} is a predefined threshold indicating the timestep at which the reward mechanism becomes active, typically set at a smaller value to suppress noise interference in the early stages of generation.

Loss function

To optimise the visual quality, artistic style, and semantic consistency of the generated murals, this study integrates latent space denoising loss[42], the conditional matching loss[43], and the stylistic loss[19,44], into a unified total loss function for the controlled generation network. The specific components are described as follows.

Latent space denoising loss is derived from the fundamental mechanism of the diffusion model, which progressively adds noise to the original mural x_0 and trains the network to reconstruct the image from this degraded representation. The objective is to preserve the structural and textural information of the image in the latent space:

$$L_{\text{denoise}} = \mathbb{E}_{q(x_t|x_0)} \left[\frac{1}{2} (\text{SNR}(t-1) - \text{SNR}(t)) \parallel x_0 - x_\theta(\alpha_t x_0 + \sigma_t \epsilon, t) \parallel^2 \right] \tag{6}$$

x_θ is the denoised mural output predicted by the model;

α_t and σ_t are time-dependent scaling parameters;

$\epsilon \sim N(0, I)$ is the random noise sampled from a standard normal distribution;

$\text{SNR}(t) = \frac{\alpha_t^2}{\sigma_t^2}$ the signal-to-noise ratio at timestep ttt, reflecting the reconstruction difficulty.

To preserve the distinctive colour palette and line aesthetics of the Kizil cave murals, a style loss is adopted to measure the similarity in stylistic features between the generated and reference images, using Gram matrices computed from intermediate layers of a pre-trained visual network:

$$L_{\text{style}} = \sum_{l=1}^{L} \lambda_l \parallel G(\phi_l(I)) - G(\phi_l(G(I, c))) \parallel_F^2 \tag{7}$$

$G(\phi_l(\cdot))$ is the Gram matrix computed from the l-th layer of the feature extractor;

$\| \cdot \|_F$ is the Frobenius norm;

λ_l is the style weight for layer l.

Finally, the total loss function is defined as a weighted combination of the individual loss terms:

$$L_{\text{total}} = \delta L_{\text{denoise}} + \eta L_{\text{condition}} + \zeta L_{\text{style}} \tag{8}$$

$$\delta = 0.4,\ \eta = 1,\ \zeta = 0.25 \tag{9}$$

Where δ, η, and ζ are the weighting coefficients that balance image quality, semantic consistency, and artistic style preservation during the restoration process.

Results

Experimental configuration

This experiment was conducted using PyTorch 2.4.0 and Python 3.11 frameworks, with training performed on four NVIDIA GeForce RTX 3090 GPUs.

Experimental comparison and analysis

To evaluate the effectiveness of the proposed method, we conducted comparative experiments against several typical image restoration models. These include Repaint, a representative diffusion-based model[28]; CTSDG, which utilises a CNN-based framework for the separation of texture and structural information[22]; ICT, a two-stage restoration method integrating Transformer and CNN architectures[23]; and SD1.5, a text-guided image restoration method[40], A dual-branch diffusion image restoration algorithm Brushnet[45] based on structure-texture decoupling modelling capabilities, and a mural restoration strategy Muralnet[46] based on line drawing-guided structure reconstruction and colour correction. To ensure a fair and rigorous evaluation of MuralNet's performance in mural restoration tasks, we standardised the structural prior sketches it relies on. Specifically, we adopted a hybrid strategy combining Sobel edge detection with manual

Table 3 | Comparison of restoration result indicators of different methods to simulate damaged murals

EvaluationMetrics	PSNR/dB				SSIM			
	M_k-1	M_k-2	M_k-3	Mean	M_k-1	M_k-2	M_k-3	Mean
Repaint	24.31	29.75	28.36	27.47	0.3765	0.5341	0.4871	0.4659
CTSDG	28.56	26.90	27.70	27.72	0.9674	0.3939	0.5250	0.6288
ICT	24.28	29.64	28.46	27.46	0.3730	0.5294	0.5198	0.4741
SD1.5	29.35	30.37	27.61	29.11	0.8312	0.7405	0.4681	0.6799
BrushNet	27.83	28.95	**29.65**	28.81	0.8477	0.3761	0.3485	0.5241
MuralNet	28.92	25.46	26.78	27.05	**0.9701**	0.3630	0.6219	0.6517
Ours	**33.51**	**31.51**	29.41	**31.48**	0.9521	**0.7613**	**0.7106**	**0.8080**

Bold text indicates the best sample results based on the evaluation metrics.

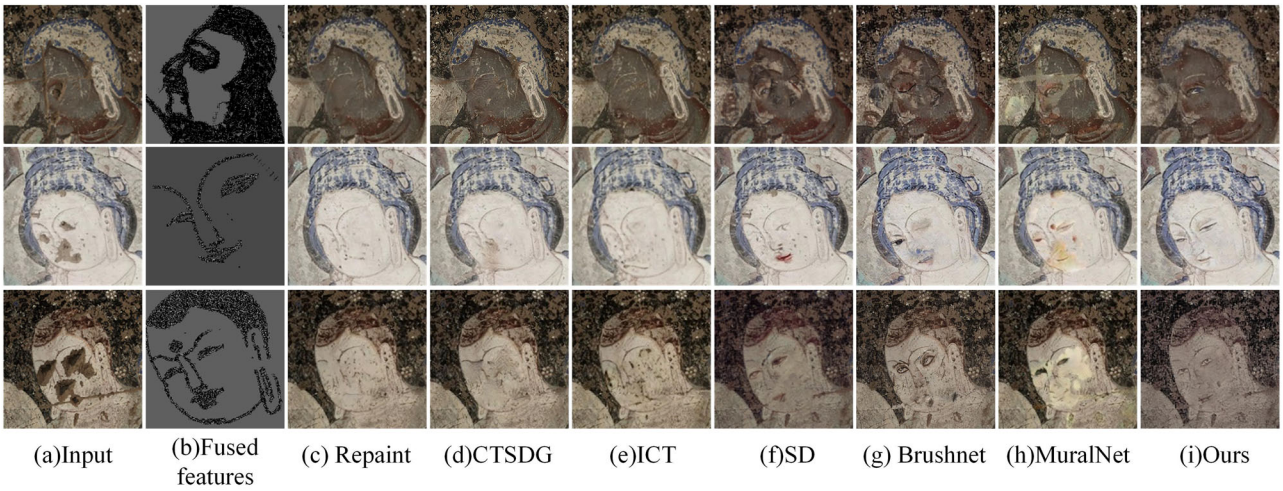

Fig. 7 | Restoration results of different algorithms on the real scene of broken mural paintings. Columns show: (**a**) Input damaged murals, (**b**) Fused features, (**c**) Repaint [28], (**d**) CTSDG [22], (**e**) ICT [23], (**f**) SD [40], (**g**) BrushNet [45], (**h**) MuralNet [46], and (**i**) Our method.

refinement. Initial edge maps were generated using the Sobel operator, followed by semantically informed manual tracing to produce clean, coherent structural line drawings. This process yielded input structure maps that are both visually complete and semantically consistent. It is important to note that edge detection algorithms often introduce structural noise or fragmented contours in damaged regions of mural images. Such artefacts can significantly impair the effectiveness of structure-guided restoration methods, particularly in areas like facial features, where restoration accuracy heavily depends on reliable priors. To minimise these confounding factors and maintain experimental consistency, all evaluations of MuralNet were conducted using manually refined structural sketches, thereby ensuring robust and noise-free guidance during the restoration process.

The effectiveness of the proposed model in mural restoration is evaluated through experiments on murals with simulated damage. To account for the irregular and continuous nature of real-world artefact degradation, masks representing facial damage ranging from 5% to 20% were designed, denoted as M_k-1–3. For MuralNet, which requires structural priors in the form of line drawings, we employed the ground-truth line drawings corresponding to the original mural images as input priors in this setting. As shown in Fig. 6, for smaller-scale mask restoration tasks, most methods are able to generate results that are nearly complete; however, issues arise in smoother regions, where the restored details often lack precision. In contrast, for larger-scale mask restoration, noticeable differences between methods emerge, with some methods struggling to maintain the consistency and structural integrity of the image content.

The restoration indices for different methods are shown in Table 3. While the PSNR[47] values for BrushNet and MuralNet in M_k-3 samples, as well as the SSIM[48] values for M_k-1 samples, are slightly higher than those of the proposed method, the overall mean values demonstrate the superiority of the proposed method in both PSNR and SSIM indices. Specifically, compared to the six methods of Repaint, CTSDG, ICT, SD, Brushnet, and Muralnet, the average PSNR scores improved by 14.60 percentage points, 13.56 percentage points, 14.64 percentage points, 8.14 percentage points, 9.27 percentage points, and 16.38 percentage points, respectively; the average SSIM scores improved by 73.43 percentage points, 28.50 percentage points, 70.43 percentage points, 18.84 percentage points, 54.17 percentage points, and 23.98 percentage points, respectively.

To further evaluate the practical effectiveness of the proposed method, we selected three representative figures from the Nirvana painting in Cave 38 of the Kizil Caves, whose facial area has been severely damaged due to religious conflict and human destruction. Figure 7 presents a comparative analysis of the restoration results under different facial mask conditions M_k-1–3 with varying occlusion ratios ranging from 5% to 20%, using different methods. As observed in Fig. 7c, d, the outputs of CTSDG and Repaint are relatively smooth but exhibit a lack of contextual semantic coherence. In Fig. 7e, the ICT method is able to reconstruct the facial structure, yet presents texture repetition, which affects the viewing experience to some extent. Figure 7f demonstrates the results of the method based on textual cues for image restoration, which provides better semantic guidance, but still has limitations in detail control. Figure 7g reveals semantic inconsistencies in the generated mural content, including stylistic deviations and misalignment in facial features, which further impact the visual authenticity of the restoration. As shown in Fig. 7h, the model exhibits noticeable blurring and pixel diffusion, which compromises the clarity and

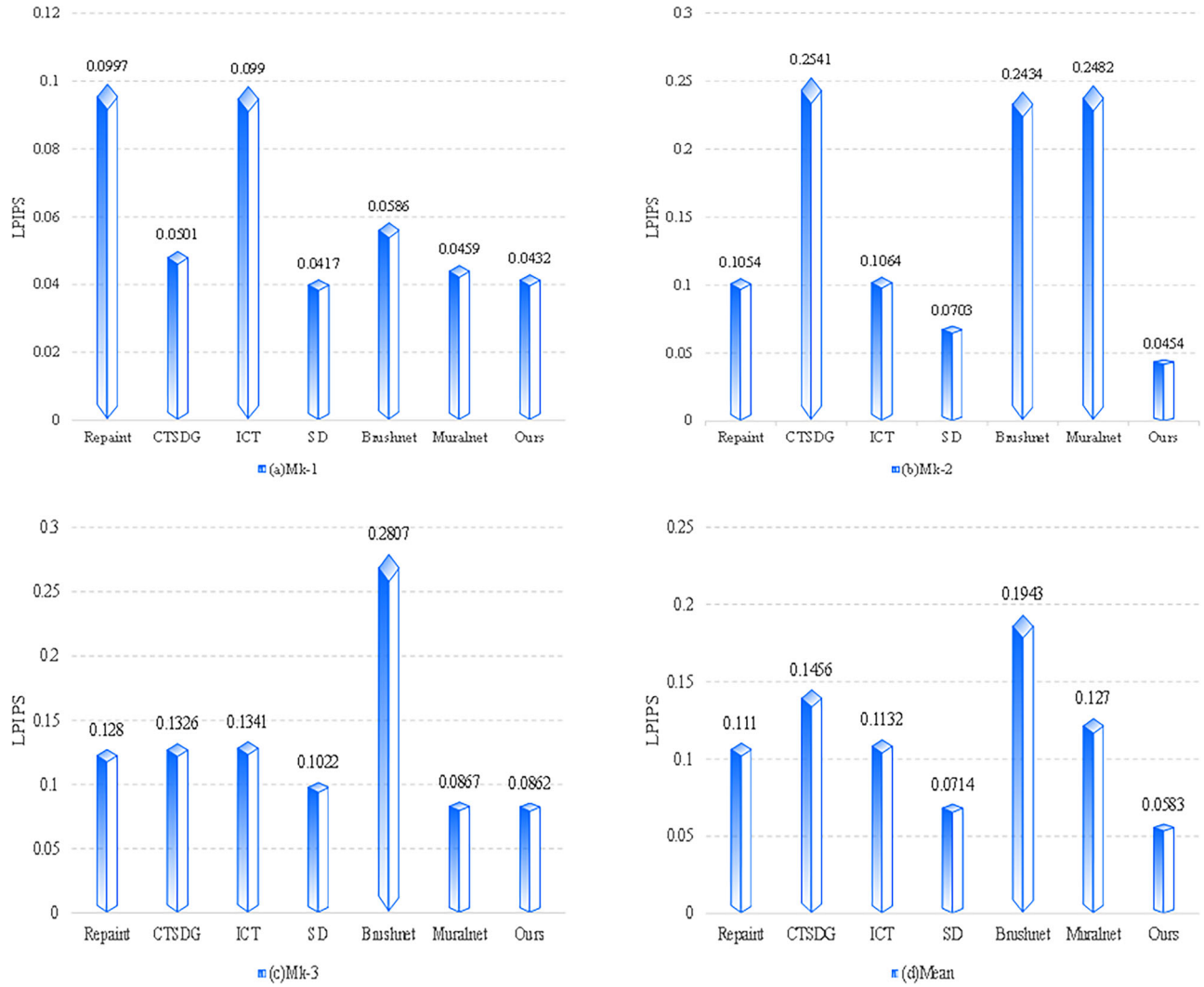

Fig. 8 | Comparison of LPIPS values for different masks. a LPIPS values of different methods under Mk-1 masks. **b** LPIPS values of different methods under Mk-2 masks. **c** LPIPS values of different methods under Mk-3 masks. **d** Average LPIPS values of different methods across all three mask types.

structural integrity of the restored region. In contrast, the method proposed in this study leverages dual conditioning—textual prompts and multi-dimensional mural feature fusion—to enable structural and semantic control. As a result, the restored images better align with the stylistic characteristics of Kizil cave murals.

Furthermore, we observed that models trained solely with conventional loss functions such as PSNR and SSIM, despite achieving high scores on these metrics, do not always produce visually optimal results. This limitation arises because the computational mechanisms of these metrics often mirror the constraints used in training, thus biasing the models toward numerical optimisation while overlooking the subtleties of human visual perception. In cultural heritage image restoration, such models may yield high-scoring results that still suffer from minor blurring in detailed areas—insufficient to meet the standards of human visual assessment.

To address this gap, we introduce LPIPS (learned perceptual image patch similarity)[49] as an auxiliary evaluation metric. LPIPS better captures perceptual differences by emulating human visual system characteristics, offering a more perceptually aligned assessment compared to PSNR and SSIM. The integration of LPIPS into our evaluation protocol aims to enhance the visual quality of the generated results—not only ensuring high scores on traditional metrics, but also satisfying perceptual demands for detail fidelity and overall realism.

As shown in Fig. 8, the LPIPS scores of the proposed method are reduced by 47.48%, 59.96%, 48.50%, 18.35%, 69.99%, and 54.09% compared to Repaint, CTSDG, ICT, SD, Brushnet, and Muralnet, respectively. In summary, the proposed method demonstrates superior performance across all three evaluation criteria—PSNR, SSIM, and LPIPS—confirming its effectiveness in both structural reconstruction and semantic coherence.

Ablation experiment

To comprehensively validate the effectiveness of each key component in the FEDR framework, we conducted five ablation experiments targeting the structural guidance path, the matching loss, and the multi-scale perception module, DFA-GSC.

As illustrated in Fig. 9, removing the structural guidance path (*w/o Structural Guidance Path*) leads to a significant decline in spatial localisation and semantic constraint capabilities for the damaged regions, resulting in distorted facial contours and blurred structures. This guidance path leverages a multi-dimensional fused structure prior and facilitates cross-layer information flow in the backbone network, substantially enhancing the model's ability to understand and reconstruct complex mural structures. When the matching loss is excluded (*w/o Matching Loss*), the overall structural layout is roughly maintained, but the generated regions exhibit stylistic inconsistencies and insufficient contextual blending. This indicates that the matching loss plays a vital role in enforcing semantic alignment between generated content and its surrounding context, thereby improving the overall restoration fidelity. We further compared the DFA-GSC module with two alternative structures—namely, the traditional GSC (*w/GSC*) and

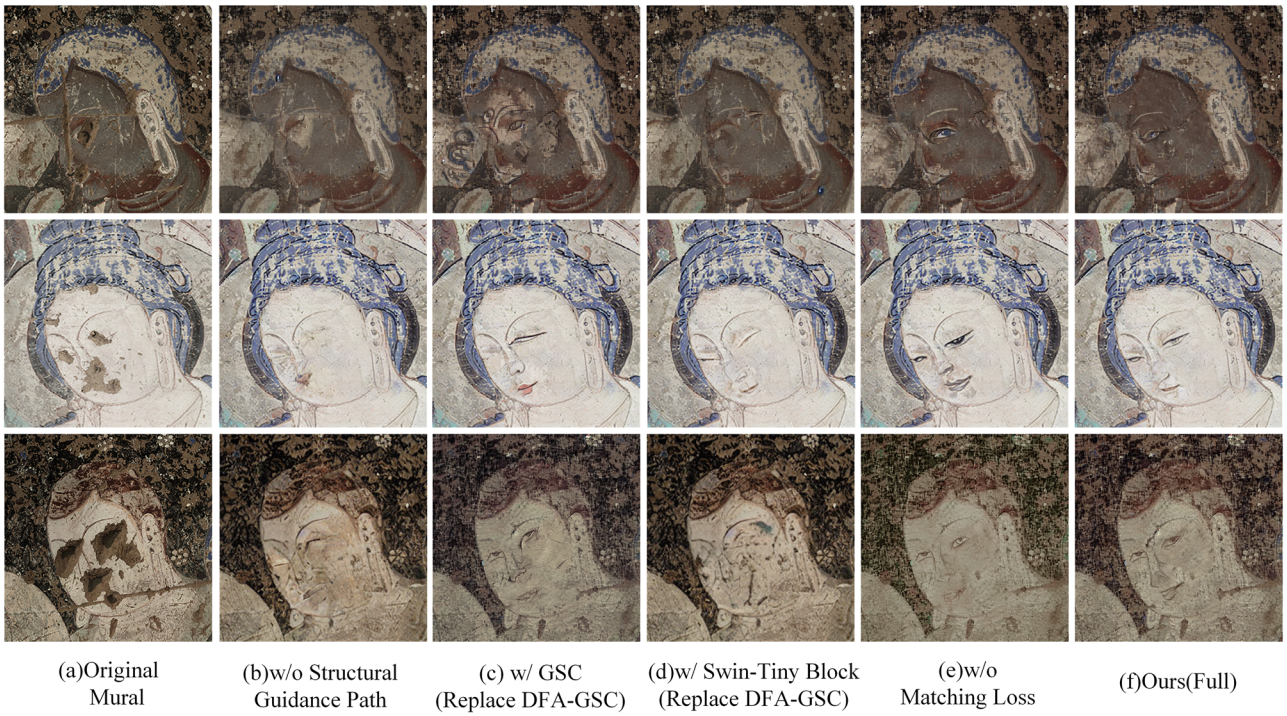

Fig. 9 | The impact of the proposed method on mural restoration quality. **a** Original mural image. **b** Restoration without the Structural Guidance Path (w/o Structural Guidance Path). **c** Restoration with DFA-GSC replaced by the traditional graph structure module GSC (w/ GSC). **d** Restoration with DFA-GSC replaced by Swin-Tiny Block (w/ Swin-Tiny Block). **e** Restoration without the Matching Loss (w/o Matching Loss). **f** Restoration using the proposed method.

Table 4 | Ablation study comparison

	w/o Structural guidance path	w/ GSC (replace DFA-GSC)	w/ Swin-tiny block (replace DFA-GSC)	w/o Matching loss	Ours (full)
PSNR↑	30.22	31.07	30.79	31.61	32.78
SSIM↑	0.7259	0.8326	0.7763	0.8561	0.8736
LPIPS↓	0.0881	0.0786	0.0803	0.0583	0.0528

the Swin Transformer Tiny Block (*w/Swin-Tiny Block*). Although both alternatives possess some modelling capacity, neither effectively captures the directional structural information that is prevalent in murals. As a result, the restored regions suffer from blurred details and unnatural transitions along semantic boundaries. In contrast, the proposed DFA-GSC module models cross-regional, direction-aware structural dependencies, leading to improved fine-grained restoration and spatial coherence.

Quantitatively, as shown in Table 4, the full model achieves superior results across all metrics: PSNR (32.78↑), SSIM (0.8736↑), and LPIPS (0.0528↓), clearly outperforming the comparative baselines. Visual comparisons further corroborate the synergistic contributions of each component in enhancing both structural restoration and texture fidelity under real-world degradation scenarios. These ablation studies confirm that the full model consistently preserves structural continuity and semantic consistency during generation, underscoring the effectiveness and necessity of the proposed method for mural restoration tasks.

Discussion

Due to their long history, natural erosion, and human-induced damage, the murals of the Kizil caves suffer from serious damage problems, posing challenges to restoration efforts. Traditional restoration methods often struggle to strike a balance between subjectivity and objectivity, thereby limiting their effectiveness in restoration.

In this study, we propose a diffusion-based restoration model framework guided by dual conditional control, incorporating both multi-feature fusion of image information and textual semantics. By assigning appropriate weights to stylistic and structural features extracted from the mural images, the model demonstrates improved capacity to recover structure and texture information during the restoration process. This contributes to enhanced consistency in the generated content, particularly in terms of texture and stylistic fidelity. To ensure semantic alignment between the input conditions and the restored output, a conditional matching loss is introduced. Additionally, the network's adaptability to complex mural features is enhanced through an improved DFA-GSC module, which strengthens the network's representational capacity.

Experimental results demonstrate that the proposed method outperforms baseline approaches in standard evaluation metrics such as PSNR, SSIM, and LPIPS, with advantages in preserving semantic consistency and detail restoration.

However, some limitations remain. The model's ability to reconstruct details is still constrained by the diversity and completeness of training data. Furthermore, in certain experimental cases, high quantitative evaluation scores are not always aligned with satisfactory visual perception. Within the context of cultural heritage restoration, there is a pressing need for domain-specific assessment metrics that more accurately reflect perceptual quality and visual coherence.

Overall, the method proposed in this paper offers an effective solution for the restoration of murals with complex textures and degraded details. Future work will focus on incorporating richer historical and cultural context into the restoration framework, through knowledge graphs and culturally guided priors, to further enhance the accuracy, interpretability, and cultural fidelity of mural restoration processes.

Data availability

All mural images used in this study were provided by the Kizil Grottoes Research Institute of the Xinjiang Uygur Autonomous Region. In accordance with the cooperation agreement between the parties, the dataset is restricted to academic research within the scope of this project. The images contain unpublished cultural relic details and sensitive mural areas. Due to cultural heritage preservation considerations and information security regulations, the data is not publicly available at this time.

Code availability

Due to institutional regulations and the sensitive nature of the project, the code cannot be shared publicly at this time.

Received: 3 April 2025; Accepted: 4 August 2025;
Published online: 31 October 2025

References

1. Ma, X. L. *Zhao Li Committee Member: Let Digital Empowerment Bring the Kizil Grottoes to Life* (Asia Center Times, 2025).
2. Chen, H. et al. DualAST: dual style-learning networks for artistic style transfer. In *Proc the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 872–881 (IEEE, 2021).
3. Haliassos, A., Barmpoutis, P., Stathaki, T., Quirke, S., Constantinides, A. Classification and detection of symbols in ancient papyri. In *Visual Computing for Cultural Heritage* (eds, Liarokapis, F., Voulodimos, A., Doulamis, N. & Doulamis, A.) 121–140 (Springer Series on Cultural Computing, 2020).
4. Bertalmio, M., Sapiro, G., Caselles, V. & Ballester, C. Image inpainting. In *Proc the 27th ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques* 417–424 (ACM Digital, 2000).
5. Efros, A. A. & Freeman, W. T. Image quilting for texture synthesis and transfer. In *Proc the 28th ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques* 341–346 (ACM Digital Library, 2001).
6. Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G. & Verdera, J. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Trans. Image Process.* **10**, 1200–1211 (2001).
7. Chen, Y., Ai, Y. & Guo, H. Improved curvature driven model for Dunhuang fresco restoration algorithm. *J. Comput. Aided Des. Comput. Graph.* **32**, 787–796 (2020).
8. Criminisi, A., Perez, P. & Toyama, K. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Process.* **13**, 1200–1212 (2004).
9. Darabi, S., Shechtman, E., Barnes, C., Goldman, D. B. & Sen, P. Image melding: combining inconsistent images using patch-based synthesis. *ACM Trans. Graph.* **31**, Article 82 (2012).
10. Hays, J. & Efros, A. A. Scene completion using millions of photographs. *Commun. ACM* **51**, 87–94 (2008).
11. Barnes, C., Shechtman, E., Finkelstein, A. & Goldman, D. PatchMatch: a randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* **28**, 24–34 (2009).
12. Shen, J., Kang, S. H. & Chan, T. F. Euler's elastica and curvature-based inpainting. *SIAM J. Appl. Math.* **63**, 564–592 (2003).
13. Lu, X. B. & Wang, W. L. Improvement of damaged Thangka image repair algorithm based on sample block. *J. Comput. Appl.* **30**, 943–946 (2010).
14. Yao, F. Damaged region filling by improved Criminisi image inpainting algorithm for Thangka. *Clust. Comput.* **22**, 13683–13691 (2019).
15. Zeng, Y., Fu, J., Chao, H. & Guo, B. Learning pyramid-context encoder network for high-quality image inpainting. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 1486–1494 (IEEE, 2019).
16. Li, J., He, F., Zhang, L., Du, B. & Tao, D. Progressive reconstruction of visual structure for image inpainting. In *2019 IEEE/CVF International Conference on Computer Vision* 5961–5970 (IEEE, 2019).
17. Suvorov, R. et al. Resolution-robust large mask inpainting with Fourier convolutions. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision* 3172–3182 (IEEE, 2021).
18. Zheng, C. et al. Bridging global context interactions for high-fidelity pluralistic image completion. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 8320–8333 (2024).
19. Nazeri, K., Ng, E., Joseph, T., Qureshi, F. Z. & Ebrahimi, M. EdgeConnect: structure guided image inpainting using edge prediction. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* 3265–3274 (IEEE, 2019).
20. Zhao, Y., Price, B., Cohen, S. & Gurari, D. Guided image inpainting: replacing an image region by pulling content from another image. In: *2019 IEEE Winter Conference on Applications of Computer Vision* 1514–1523 (IEEE, 2019).
21. Zhang, L., Chen, Q., Hu, B. & Jiang, S. Text-guided neural image inpainting. In *Proc of the 28th ACM International Conference on Multimedia* 1302–1310 (IEEE, 2020).
22. Guo, X., Yang, H. & Huang, D. Image inpainting via conditional texture and structure dual generation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 14114–14123 (IEEE, 2021).
23. Wan, Z., Zhang, J., Chen, D. & Liao, J. High-fidelity pluralistic image completion with transformers. In *2021 IEEE/CVF International Conference on Computer Vision* 4672–4681 (IEEE, 2021).
24. Xu, W. & Fu, Y. Deep learning algorithm in ancient relics image colour restoration technology. *Multimed. Tools Appl.* **82**, 23119–23150 (2023).
25. Peng, X. et al. C3N: content-constrained convolutional network for mural image completion. *Neural Comput. Appl.* **35**, 1959–1970 (2023).
26. Yang, J., Ruhaiyem, N. I. & Zhou, C. A 3M-hybrid model for the restoration of unique giant murals: a case study on the murals of Yongle Palace. *IEEE Access* **13**, 38809–38824 (2025).
27. Wang, F., Wu, K. & Cao, Y. High resolution mural images inpainting based on dual aggregation generative adversarial network. In *2023 International Conference on Image Processing, Computer Vision and Machine Learning* 97–102 (IEEE, 2023).
28. Wang, J. et al. Image editing with diffusion models: a survey. https://arxiv.org/abs/2504.13226 (2025).
29. Lugmayr, A. et al. RePaint: inpainting using denoising diffusion probabilistic models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 11451–11461 (IEEE, 2022).
30. Meng, C. et al. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *The Tenth International Conference on Learning Representations* (ICLR, 2022).
31. Wang, Y., Yu, J. & Zhang, J. Zero-shot image restoration using denoising diffusion null-space model. In *The Eleventh International Conference on Learning Representations* (ICLR, 2023).
32. Xia, B. et al. (2023). DiffIR: efficient diffusion model for image restoration. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 13049–13059 (IEEE, 2023).
33. Zhang, L. et al. Minutes to seconds: speeded-up DDPM-based image inpainting with coarse-to-fine sampling. In *2024 IEEE International Conference on Multimedia and Expo (ICME)* 1–6 (IEEE, 2024).
34. Saharia, C. et al. Palette: image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, Article 15 (ACM, 2022).
35. Huang, Y. et al. WaveDM: wavelet-based diffusion models for image restoration. *IEEE Trans. Multimed.* **26**, 7058–7073 (2023).
36. Huang, J. et al. Dual-schedule inversion: training- and tuning-free inversion for real image editing. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision* 660–669 (IEEE, 2024).
37. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)* 8748–8763 (PMLR, 2021).
38. Zhang, L., Rao, A. & Agrawala, M. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 3813–3824 (IEEE, 2023).

39. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. https://arxiv.org/abs/1312.6114 (2013).
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10684–10695 (IEEE, 2022).
41. Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. CBAM: Convolutional Block Attention Module. In *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, vol. 11211, 3–19 (Springer, Cham, 2018).
42. Ho, J., Jain, A. & Abbeel, P. Denoising Diffusion Probabilistic Models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 574–585 (Curran Associates Inc., Red Hook, NY, 2020).
43. Li, M. et al. ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback. In *European Conference on Computer Vision* (ECCV, 2024).
44. Yang, S., Jiang, L., Liu, Z. & Loy, C. C. Pastiche Master: Exemplar-Based High-Resolution Portrait Style Transfer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7683–7692 (IEEE, 2022).
45. Ju, X. et al. BrushNet: A Plug-and-Play Image Inpainting Model with Decomposed Dual-Branch Diffusion. In *Computer Vision – ECCV* 2024, 150–168 (2024).
46. Li, L. et al. Line Drawing Guided Progressive Inpainting for Mural Damage. *J. Comput. Cult. Herit.* (2025).
47. Gupta, P., Srivastava, P., Bhardwaj, S. & Bhateja, V. A modified PSNR metric based on HVS for quality assessment of color images. In *Proc the 2011 International Conference on Communication and Industrial Application* 1–4 (ICCIA, 2011).
48. Horé, A. & Ziou, D. Image quality metrics: PSNR vs. SSIM. In *Proc the 2010 20th International Conference on Pattern Recognition* 2366–2369 (ICPR, 2010).
49. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595 (IEEE, 2018).

Acknowledgements

This work was supported by the Key Research and Development Special Project of the Xinjiang Uygur Autonomous Region: “Digital Restoration and Immersive Experience Demonstration of Key Cave Murals in the Kizil Caves” (project no. 2022397539).

Author contributions

L.G.Y. contributed to the conceptualization, methodology, and data curation of the study. W.H.Q. and W.K. participated in the conceptualization and contributed to the methodology. L.G.Y. prepared the original draft of the manuscript. L.G.Y., W.H.Q., W.K., Z.H.D., and Z.L. were responsible for reviewing and editing the manuscript. Z.L. acquired the funding for the project. All authors have read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Huiqin Wang.