npj | heritage science

Article

https://doi.org/10.1038/s40494-026-02325-w

M3SFormer: multi-stage semantic and style-fused transformer for mural image inpainting

Check for updates

Qiyao Hu[1,2], Qinfan Ge[1], Yihan Zhang[1], Xianlin Peng[3,4] ✉, Jiangpeng Wang[3], Shuyi Qu[1,5] & Nana Chen[6] ✉

Digital restoration of ancient murals is crucial for preserving cultural heritage. However, existing methods often suffer from semantic distortion and stylistic inconsistencies when repairing large damaged areas. This paper proposes M3SFormer, an innovative restoration framework. Built on an enhanced P-VQVAE module, it employs continuous feature modeling without quantization to retain details. A new Semantic-Style Consistency Module (SSCM) integrates regional semantics with multi-scale style features, ensuring coherent outputs. Furthermore, the Flow-Guided Refinement Module (FGRM) reconstructs key textures through network guidance, improving visual quality. Experiments on multiple benchmarks show that M3SFormer surpasses mainstream methods across all metrics, with significant gains in PSNR, SSIM, and LPIPS. It also excels in reconstructing complex structures and preserving styles under high mask coverage, offering a reliable solution for high-quality digital mural preservation. The dataset and code are available at: https://github.com/LPDLG/M3SFormer

Ancient murals stand as vital material cultural heritage of Chinese civilization. They not only trace the developmental trajectory of ancient Chinese art but also bear rich historical, religious, and socio-cultural information[1]. These murals demonstrate exceptional artistic mastery in their compositional forms, color applications, and expressive techniques. Their visual language reflects the esthetic ideals and philosophical concepts of specific historical periods, holding immense historical and artistic value[2]. However, due to the long-term effects of natural weathering, human damage, and other factors, many murals have suffered severe deterioration, including pigment flaking, structural fractures, and blurred imagery. This not only results in the loss of visual information but also undermines the original cultural expression and historical authenticity of the murals, necessitating urgent effective inpainting and digital preservation through modern technological means[3].

With the continuous advancement of deep learning in the field of computer vision, image inpainting techniques have become one of the key supporting tools for the digital preservation of murals[4]. Early methods based on convolutional neural networks (CNNs), such as EdgeConnect[5], relied on local receptive fields to process missing regions in images. However, when dealing with large-area defects, these approaches often produced blurry or structurally discontinuous results. Some image processing ideas introduced in lightweight CNN architectures[6–8], as well as concepts from CNNs designed for object detection and classification[9–11], have also provided inspirational technical support for mural restoration. Furthermore, the application of lightweight object detection methods to mural images[12] has shown positive implications for the inpainting task. With the introduction of generative adversarial networks (GANs), certain progress has been made in texture generation for image inpainting. Nevertheless, these methods still face challenges in ensuring the authenticity and historical accuracy of the restored results.

Overall, existing methods face numerous challenges when addressing mural restoration tasks. Mural images typically feature complex structures and distinct stylistic characteristics, making it difficult for traditional restoration methods to simultaneously meet the high demands for detail fidelity and structural consistency.

• Traditional methods based on CNNs perform well in texture filling and local structure reconstruction. However, due to their limited receptive fields, they often struggle to model long-range dependencies. This limitation is particularly evident when handling large-scale missing areas or complex semantic scenes, where issues such as semantic misalignment and structural discontinuities frequently arise.

[1]School of Electronic Information, Northwest University, Shannxi, China. [2]State-Province Joint Engineering and Research Center of Advanced Networking and Intelligent Information Services, Xi'an, China. [3]School of Art, Northwest University, Xi'an, China. [4]Shaanxi Silk Road Cultural Heritage Digital Protection and Inheritance Collaborative Innovation Center, Xi'an, China. [5]Shaanxi Key Laboratory of Higher Education Institution of Generative Artificial Intelligence and Mixed Reality, Xi'an, China. [6]Key Laboratory of Archaeological Exploration and Cultural Heritage Conservation Technology, Ministry of Education, Northwestern Polytechnical University, Xi'an, China. ✉e-mail: pxl@nwu.edu.cn; chen_nana@nwpu.edu.cn

• Meanwhile, most methods rely on discrete quantization strategies during image encoding. While this compression mechanism helps reduce computational complexity, it inevitably leads to the loss of high-frequency information, thereby weakening the ability to restore intricate brushstroke details. On the other hand, current approaches generally exhibit weak structural guidance in the inpainting process, lacking effective modeling of local detail variation trends.

• Moreover, stylistic consistency, a critical yet often overlooked dimension in mural inpainting, is frequently neglected due to the absence of stylistic constraints. This leads to inconsistencies in color and other aspects between inpainting areas and the original image, thereby compromising the overall visual unity.

• Finally, stylistic authenticity and structural accuracy are fundamental requirements for mural restoration. Any structural reconstruction that fails to harmonize with the surrounding environment or stylistic reproduction that lacks authenticity will result in severe visual discord between the restored work and the original mural, thereby causing irreversible damage to its overall historical and artistic value.

To address the aforementioned challenges, we propose **M3SFormer, M**ulti-**S**tage **S**emantic and **S**tyle-Fused Trans**former** for Mural Image Inpainting. The framework employs a coarse-to-fine multi-stage inpainting workflow and introduces a guidance mechanism based on fine-grained flow field prediction. By estimating local structural transformation trends, it directs the network to focus on critical detail regions, achieving simultaneous enhancement of semantic consistency and structural stability. The main contributions of this work are as follows:

• We propose the Global Structure Reasoning Module (GSRF), which first introduces a continuous feature modeling strategy. Abandoning traditional discrete quantization schemes, it employs an improved P-VQVAE encoder to perform quantization-free modeling of images. This approach preserves richer detail features and texture information, thereby enhancing the inpainting quality of high-frequency structures.

• We propose the Semantic-Stylistic Consistency Module (SSCM), which leverages regional semantic information from the SMTs network and multi-level perceptual style features extracted by the VGG network. This module is jointly optimized through guided loss and Gram matrix style loss. To further enhance the stylistic alignment between the repaired region and the original image, we integrated dual prior information from semantic segmentation and style matching.

• We propose the Flow-Guided Refinement Module (FGRM), which employs a flow-regularized dynamic optimization framework. By modeling the refinement process as a continuous state evolution system, it enables fine-grained adjustments to refinement outcomes.

M3SFormer aims to achieve structural precision, semantic clarity, and stylistic consistency through comprehensive optimization across the entire workflow, from feature modeling and structural guidance to the construction of multi-dimensional loss functions. Experimental validation demonstrates that this method outperforms current mainstream inpainting algorithms across multiple datasets, exhibiting particularly significant advantages in restoring intricate details within complex regions and preserving artistic styles. It provides a more robust and scalable technical approach for the digital inpainting of ancient murals.

• Transformer-based Image inpainting

In recent years, deep learning has gained significant traction in image inpainting. The Transformer architecture, with its robust capability for modeling long-range dependencies, has achieved breakthrough progress in image inpainting tasks. Vision Transformers (ViTs) effectively extract global information by processing images in chunks. However, standard ViTs face computational complexity that scales quadratically with image size when handling high-resolution images. To address this, Swin Transformers strike a balance between global modeling capabilities and computational efficiency by introducing a hierarchical window attention mechanism.

However, existing Transformer-based restoration methods still face numerous challenges in mural scenarios, particularly exhibiting suboptimal performance in detail recovery and style fidelity. Although autoregressive models like VQ-Transformer[13] have made progress in image generation quality, their feature discretization and quantization processes are prone to high-frequency information loss, thereby weakening their ability to express subtle characteristics such as mural textures and brushstrokes. On the other hand, methods like IM-CTSDG[14] enhance pigment texture restoration through multi-scale contextual feature fusion. However, their backbone networks remain convolutional-based, struggling to model long-range dependencies. Consequently, when handling large-scale structural gaps or misalignments common in murals, they often produce semantic and geometric inconsistencies between restored regions and surrounding structures, leading to visual disharmony. Furthermore, color transitions between structural regions in murals are often more pronounced than in natural images, amplifying the issues caused by structural inconsistencies. To address these challenges, this study adopts an improved UQ-Transformer architecture[15], which introduces continuous feature representations to mitigate quantization information loss. Combined with a multi-token sampling strategy to enhance inference efficiency, this approach achieves a better balance between overall structural consistency and local detail restoration quality.

Additionally, PUT incorporates an Attention State Space Module (ASSM)[16], when combined with a semantic normalization mechanism, effectively models repetitive patterns and large-scale structures within murals. This enhances structural continuity and semantic consistency. It demonstrates superior performance compared to traditional convolutional models, particularly in reconstructing complex materials and regional boundaries[17].

• Multi-stage inpainting

In recent years, image inpainting methods based on diffusion models have demonstrated excellent generation quality, particularly in complex damage scenarios. The RePaint model[18] generates plausible inpainting content through progressive denoising. However, such methods typically require extensive iterative computations, resulting in low inference efficiency. To enhance inpainting efficiency and effectiveness, Rectified Flow[19] introduces an ordinary differential equation (ODE) optimization path. This approach significantly improves computational efficiency while maintaining inpainting quality[20].

The multi-stage mechanism was also adopted in Efficient Diffusion[21], which significantly reduced the number of iterations through residual transfer, thereby achieving a better trade-off between repair accuracy and efficiency. Furthermore, the study enhanced the mechanism's information transfer during the repair process by introducing an adaptive loss function and dynamic weight adjustment based on semantic labels[22].

However, existing methods often suffer from flaws due to excessive hallucination: the restored area may appear structurally incongruous with its surroundings, or when handling abrupt color changes, it may generate unnatural smooth transitions instead of preserving the necessary visual contrast.

Overall, multi-stage restoration methods excel at handling images with large-scale defects and complex structures, such as murals. Their core strength lies in the progressive restoration process from structure to texture. However, existing approaches still face limitations. To address this, we incorporate the open-set semantic segmentation capability of Mask2Former[23], enabling it to more flexibly identify and process special elements within murals. This significantly enhances the model's adaptability to complex styles and content.

• Style consistency inpainting

Maintaining stylistic consistency is crucial in mural inpainting. Existing style preservation methods predominantly rely on Gram matrix-based style loss functions, which struggle to capture local variations in detail within murals, particularly when different materials are used across distinct areas of the artwork.

Specifically, by combining Mask2Former[23] semantic guidance capability with style features extracted by the VGG network, we can

independently model style within different semantic regions, ensuring stylistic consistency and coherence of local features.

Meanwhile,[24] emphasizes the incorporation of style distribution learning in mural inpainting to enhance coherence between areas of different materials. Our semantic grouping style loss method partitions images based on Mask2Former[23] outputs and calculates independent Gram matrix differences for each region, thereby achieving more targeted style representation. Some studies[25] have also attempted to couple style preservation with the inpainting workflow into a closed-loop system, such as by applying style residual feedback at each generation step.

Methods

Ancient murals, with their diverse styles, constitute a vital component of cultural heritage. Their preservation and inpainting hold significant importance for cultural continuity. We propose an incremental mural inpainting framework that progressively recovers the original appearance of damaged murals through multi-stage, meticulous treatment.

First, we designed the Global Structural Reasoning Module (GSRF), which effectively models long-range dependencies in images by incorporating continuous feature encoding and self-attention mechanisms, thereby avoiding information loss issues in global inpainting that plague traditional methods. Second, we introduced the semantic stylistic consistency module (SSCM). This module employs a semantic mapping transformer (SMT) and a hierarchical semantic-aware style loss to ensure inpainting results maintain high consistency with the original mural in both semantic structure and artistic style. Finally, we developed the FGRM. This module employs a flow-regularized dynamic optimization framework, modeling the inpainting process as a continuous state evolution system. Through iterative optimization, it progressively enhances inpainting quality. This approach not only effectively handles murals with varying degrees of damage but also preserves the characteristics of different artistic styles (Fig. 1).

Overall structure

The mural inpainting framework proposed in this study effectively enhances inpainting outcomes through the synergistic interaction of multiple modules. As illustrated in Fig. 2.

We demonstrate the complete inpainting workflow of M3SFormer, featuring a hierarchical organization and information exchange among key components, including feature extraction, semantic understanding, style constraints, and flow optimization. Ensuring efficient execution of inpainting tasks across multiple levels.

GSRF Module is the core component of the global structural reasoning module, the AQ-Transformer module employs continuous feature encoding and self-attention mechanisms to effectively model long-range dependencies within images. Through an enhanced attention mechanism and dynamic position encoding, this module avoids the information loss issues inherent in traditional methods during global inpainting, providing high-quality feature representations for subsequent processing stages.

SSCM Module integrates semantic segmentation with style matching mechanisms, guiding inpainting through the extraction of high-level semantic features. Its hierarchical semantic awareness mechanism ensures precise recovery of style characteristics across distinct semantic regions. It comprises the SLSM module and SMT module group.

SLCM Module is a multi-level feature extraction scheme based on the VGG network precisely controls the artistic style consistency of the inpainting area. This module receives semantic segmentation results and image features, calculates hierarchical Gram matrices, and optimizes the semantic-aware style loss to ensure the inpainting area maintains high consistency with the original mural in artistic characteristics such as brushstrokes and color.

SMT Module Group consists of three SMTs that enhance the semantic perception capabilities through a parallel semantic feature extraction mechanism. Each SMT module independently processes input features and outputs semantic information, ultimately fusing to generate precise semantic segmentation maps. These maps provide reliable semantic guidance for subsequent style constraint and flow optimization tasks.

FGRM Module employs a flow-regularized dynamic optimization framework, refining inpainting outcomes by modeling the inpainting process as a continuous state evolution system. This module progressively enhances inpainting quality through iterative optimization guided by semantic information, ensuring high precision and fidelity throughout the inpainting process.

Global structure reasoning module

The global structural reasoning module (GSRF) successfully achieves global structural modeling of images through image block feature extraction, self-attention mechanisms, global dependency modeling, and the introduction of positional encoding. This module not only preserves local image features but also enhances the model perception of global structure through techniques like long-range dependency modeling and positional encoding, providing a solid foundation for subsequent inpainting modules. Through multi-head attention mechanisms and meticulous spatial information modeling, the GSRF module effectively captures complex structural features within images, thereby improving the accuracy and artistic quality of inpainting images. Figure 3 shows the UQ-Transformer before improvement and the improved AQ-Transformer.

During mural inpainting, global structural modeling is crucial for ensuring the inpainting results preserve the original image information. To overcome the limitations of traditional methods in local inpainting, this study employs a Transformer-based spatial position encoding approach. This method utilizes a multi-head self-attention mechanism to effectively model long-range dependencies within the image, thereby preventing information loss and achieving accurate inpainting of global information. To assist the model in understanding the relative spatial positions of image blocks, we introduce a two-dimensional sine position encoding.

Image block feature extraction. To capture global features of an image, we partition the input image $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ into multiple image patches $\mathbf{p}_i \in \mathbb{R}^{r\times r\times 3}$ of size $r \times r \times 3$. Each patch undergoes a linear transformation to map it into a high-dimensional feature space. The feature vector $\mathbf{F}_i \in \mathbb{R}^D$ for each block is computed as follows:

$$\mathbf{F}_i = \sum_{k=1}^{r^2\cdot 3} \mathbf{W}_{p,k} \cdot \mathbf{p}_{i,k} + \mathbf{b}_p = \sum_{k=1}^{r^2\cdot 3} \mathbf{W}_{p,k} \cdot \left(vec(\mathbf{p}_i)\right)_k + \mathbf{b}_p \tag{1}$$

where $\mathbf{W}_p \in \mathbb{R}^{D\times(r^2\cdot 3)}$ is the weight matrix for the linear transformation, $\mathbf{b}_p$ is the bias term, and $vec(\mathbf{p}_i)$ denotes the vectorization operation applied to the image patch $\mathbf{p}_i$.

In this manner, features from all image patches are combined into a feature matrix $\mathbf{F} \in \mathbb{R}^{N\times D}$, where N is the total number of image patches, $D = 768$ is the feature dimension, and $r = 16$ is the patch size.

This feature extraction process not only preserves local details but also provides rich feature information for subsequent global modeling and inpainting tasks. This approach ensures that each local region of the input image is transformed into a highly expressive feature vector, enabling efficient information exchange and fusion through the self-attention mechanism in subsequent stages.

To capture global dependencies between image blocks, we introduce a self-attention mechanism. Given an input feature matrix $\mathbf{F} \in \mathbb{R}^{N\times D}$, we employ a query matrix $\mathbf{Q} \in \mathbb{R}^{N\times D}$, key matrix $\mathbf{K} \in \mathbb{R}^{N\times D}$, and value matrix $\mathbf{V} \in \mathbb{R}^{N\times D}$. The output feature matrix is obtained through weighted summation:

$$\mathbf{F}_o = \text{softmax}\left(\frac{\mathbf{Q}\cdot\mathbf{K}^T}{\sqrt{D}}\right)\cdot\mathbf{V} \tag{2}$$

This process calculates weights through query-key similarity and performs weighted aggregation of numerical vectors, thereby effectively

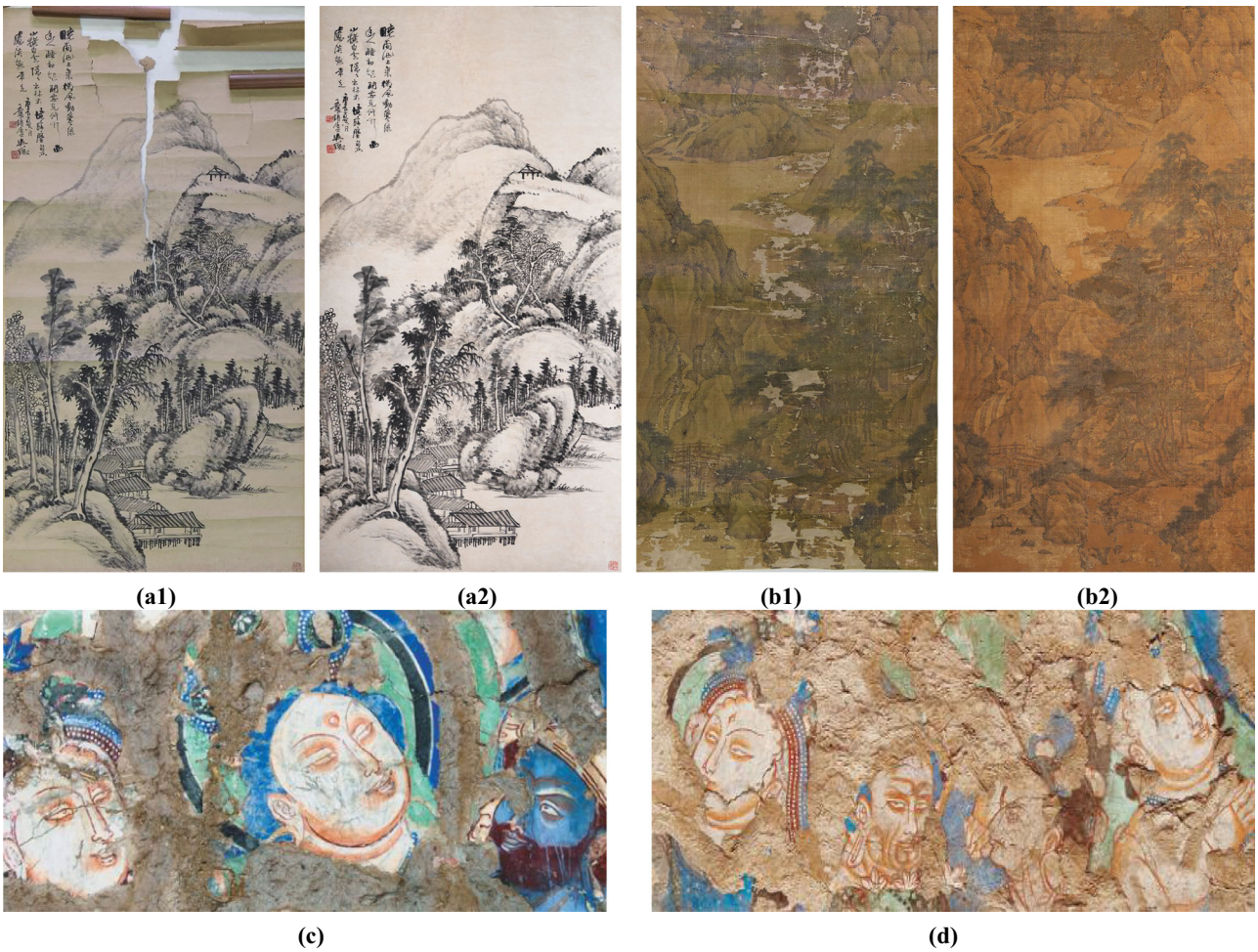


Fig. 1 | Illustration of the damaged murals and landscape paintings. a1, **a2** show pre-inpainting and AI-inpainting ink-wash landscapes by modern artist *Daiqiu Wu*. **b1**, **b2** depict a colored landscape with figures by *Ming* Dynasty artist *Hong Yang*. **c**, **d** display genuinely damaged, uninpainting murals.

modeling global relationships between image blocks and enhancing context-aware capabilities during the inpainting process.

The multi-head attention mechanism enables the model to learn features in parallel across multiple subspaces, thereby better capturing different dependencies and details within images. Using the multi-head attention mechanism, the output feature representation is:

$$\mathbf{F}_t = Concat(\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_h) \cdot \mathbf{W}_O \tag{3}$$

Among these, $\mathbf{W}_O \in \mathbb{R}^{hD_v \times D}$ is the output weight matrix, $\mathbf{F}_i$ is the output of each attention head, $h = 12$ is the number of heads, and D_v is the output dimension of each head.

Spatial position encoding. To help the model understand the relative spatial positions of image patches, we introduce a two-dimensional sine position encoding. This encoding incorporates spatial information into the model by encoding the positional indices of image patches, thereby enhancing the model's understanding of spatial structures. The formula for the two-dimensional sine position encoding is as follows:

$$\mathbf{P}_{i,j,2d} = \sin\left(\frac{i}{10000^{2d/D}}\right) \tag{4}$$

$$\mathbf{P}_{i,j,2d+1} = \cos\left(\frac{j}{10000^{2d/D}}\right) \tag{5}$$

where D denotes the feature dimension, while i and j represent the row and column position indices of the image block within the image, respectively. Position encoding is added to the input features to ensure that spatial location information is preserved.

This two-dimensional position encoding method aligns with the standard practice in Vision Transformers (ViT). By controlling wavelength decay $10000^{2d/D}$, it endows encodings at different positions with distinct frequency characteristics, thereby enabling the model to better capture spatial information.

Semantic-stylistic consistency module

To ensure inpainting results maintain high consistency with the original mural in both semantic structure and artistic style, we propose the SSCM. This module serves a dual guiding role within the inpainting framework: on one hand, it provides semantic prior information to the subsequent FGRM, guiding structural generation during inpainting. On the other hand, it employs a hierarchical semantic perception mechanism to ensure precise recovery of stylistic features across different semantic regions.

Semantic Segmentation Guidance. In image inpainting tasks, accurate semantic understanding is crucial for ensuring the plausibility of inpainting results. We extract high-level semantic features through a set of parallel SMTs. Each SMT module consists of a multi-head self-attention mechanism and a feedforward network, which generate precise semantic vectors through deep processing of input features.

Finally, the softmax fusion yields a semantic label matrix $\mathbf{S} \in \mathbb{R}^{N \times M}$, where N denotes the number of image patches and M represents the number of semantic categories. The semantic-guided feature reconstruction process

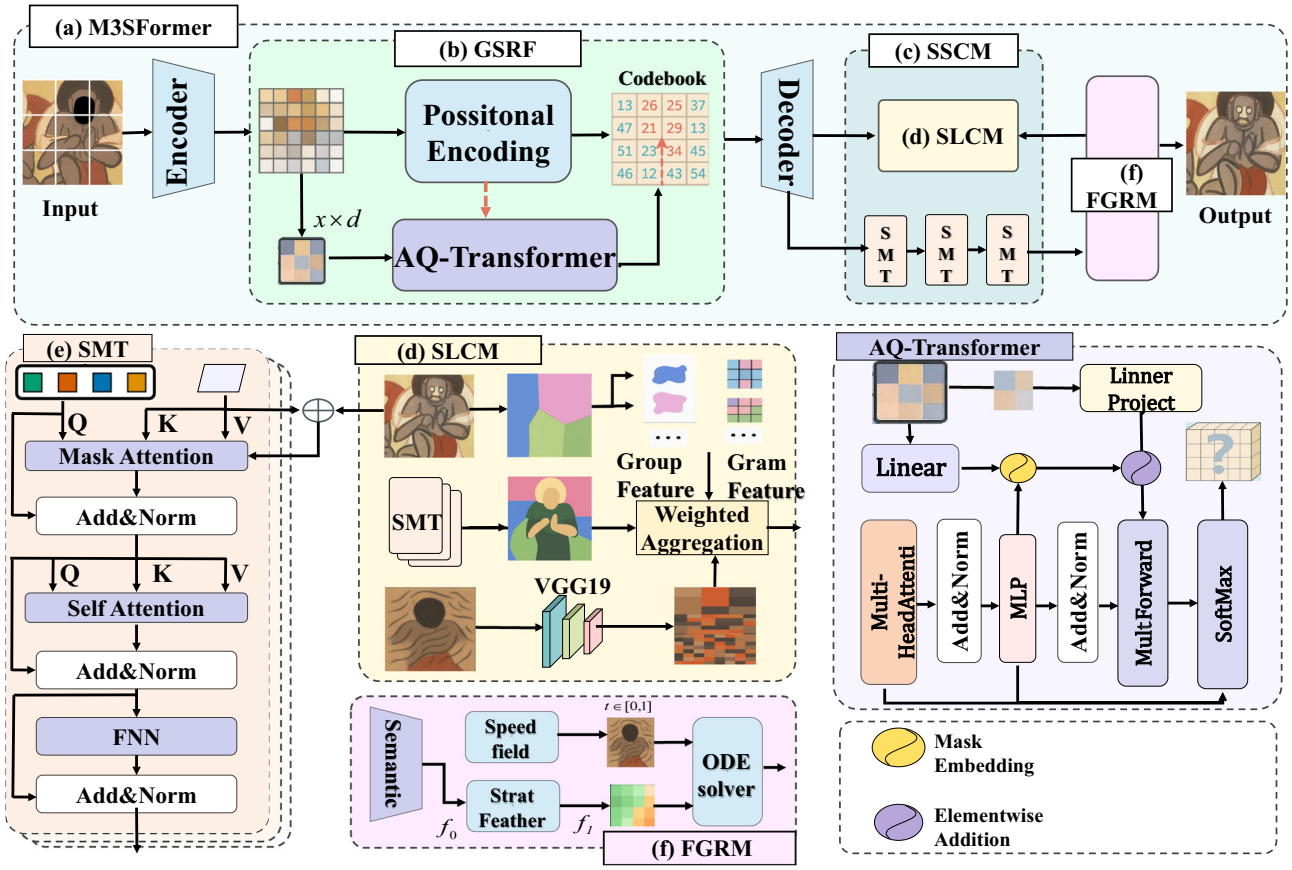
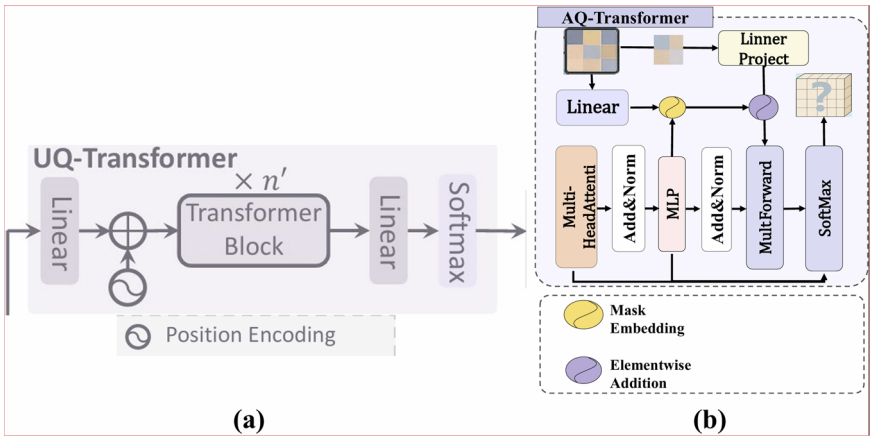


Fig. 2 | Illustration of the M3SFormer. a M3SFormer **b** GSRF **c** SSCM **d** SLCM **e** SMT **f** FGRM.

Fig. 3 | UQ-Transformer before and after improvement **a** UQ-Transformer **b** AQ-Transformer.



is represented as:

$$\mathbf{F}_{guide} = \sum_{i=1}^{N} \mathbf{S}_{i,j} \cdot \mathbf{F}_{opt,i} \tag{6}$$

This semantic information serves a dual purpose: First, it provides structured prior knowledge to the FGRM module, guiding the flow optimization process through the semantic mask $\mathbf{M}_c$ to evolve within specific semantic regions, ensuring the inpainting process adheres to semantic constraints. Second, it offers grouping criteria for style loss computation, enabling independent style matching across different semantic regions. The semantic segmentation loss function is defined as:

$$\mathcal{L}_{sem} = -\sum_{c=1}^{C} y_c \log(p_c) \tag{7}$$

This loss ensures the network can accurately identify various semantic regions, laying the foundation for subsequent semantic-guided inpainting.

SLCM. Based on semantic segmentation results, we propose a hierarchical semantic-aware style matching mechanism. This mechanism employs a pre-trained VGG19 network as the feature extractor, extracting

feature representations at three levels: {*relu*1_2, *relu*2_2, *relu*3_3}. For each semantic category c, we first extract the feature map of the corresponding region using the semantic mask $\mathbf{M}_c$:

$$\mathbf{F}_{out}^{c} = \mathbf{M}_c \odot \mathbf{F}_{out} \tag{8}$$

$$\mathbf{F}_{gt}^{c} = \mathbf{M}_c \odot \mathbf{F}_{gt} \tag{9}$$

Then compute the Gram matrix for each semantic region at different layers:

$$\mathbf{G}_l^c(\mathbf{F}_{out}) = \mathbf{F}_{out}^{c,l} \cdot (\mathbf{F}_{out}^{c,l})^\top \tag{10}$$

$$\mathbf{G}_l^c(\mathbf{F}_{gt}) = \mathbf{F}_{gt}^{c,l} \cdot (\mathbf{F}_{gt}^{c,l})^\top \tag{11}$$

The layered semantic-aware style loss is ultimately defined as:

$$\mathcal{L}_{style} = \sum_{c=1}^{C} \sum_{l \in \mathcal{L}} \lambda_l \left\| \mathbf{G}_l^c(\mathbf{F}_{out}) - \mathbf{G}_l^c(\mathbf{F}_{gt}) \right\|_F^2 \tag{12}$$

where $\mathcal{L}$ = {*relu*1_2, *relu*2_2, *relu*3_3} denotes the selected feature layer set, and λ_l represents the layer weight coefficient. This design ensures that different semantic regions, such as facial features, clothing textures, and background decorations, can undergo independent style matching, thereby avoiding feature smoothing issues that might arise from global style loss.

Flow-guided refinement module

Traditional one-shot forward generation methods often struggle with complex structural damage and semantic inconsistencies. To address this, we propose a flow-regularized progressive optimization framework that models the inpainting process as a continuous state evolution system. The core idea is to gradually evolve the initial inpainting result toward the final output along a semantically guided optimization path by learning a parameterized velocity field.

Specifically, given the initial repair result X_0 generated by the PUT network and the target image X_1, we construct a continuous time series $\{X_t\}_{t\in[0,1]}$, whose evolution is described by the following ODE:

$$\frac{dX_t}{dt} = v_\theta(X_t, t|\mathcal{M}) \tag{13}$$

where v_θ denotes the learnable velocity field network, and $\mathcal{M}$ represents the semantic guidance information provided by Mask2Former. The velocity field network adopts a U-Net architecture, with its input being the concatenation of the current state X_t and semantic feature maps, and its output being a three-dimensional flow field.

In terms of implementation details, we employ the Euler method with fixed step size for numerical solution:

$$X_{t+\Delta t} = X_t + \Delta t \cdot v_\theta(X_t, t|\mathcal{M}) \tag{14}$$

The step size Δt is set to 0.1, with a total of 10 iterations performed. This design ensures repair quality while keeping computational overhead within acceptable limits. Notably, this framework is fully decoupled from the style constraint module to be introduced later. This modular design enables independent optimization and upgrades of each component.

Loss function

This loss function design embodies tight coupling between modules: $\mathcal{L}_{sem}$ provides regional guidance for style matching while imposing structural constraints for flow optimization. The style loss $\mathcal{L}_{style}$ ensures artistic authenticity in the inpainting results. and the flow optimization loss $\mathcal{L}_{flow}$ performs the final refinement under dual constraints of semantics and style.

Through this collaborative optimization mechanism, our framework maintains semantic plausibility and stylistic consistency in repair results even under complex damage conditions. Based on the aforementioned module design, we constructed a multi-level total loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{flow} + \lambda_1 \mathcal{L}_{sem} + \lambda_2 \mathcal{L}_{style} \tag{15}$$

where $\mathcal{L}_{flow}$ is the flow optimization loss, ensuring distribution alignment during the inpainting process. $\mathcal{L}_{sem}$ is the semantic segmentation loss, guaranteeing semantic consistency. $\mathcal{L}_{style}$ is the hierarchical, semantic-aware style loss, maintaining artistic style consistency. Hyperparameters λ_1 and λ_2 are used to balance the contribution of each loss term.

Results

Datasets

We utilize two datasets, MuralVerse-S and MaskCLP-S, in the experiments to assess the effectiveness of the proposed method.

We propose a dataset MuralVerse-S of murals compiled from publicly available images across various regions of China. It comprises 1396 extended and cropped images of Dunhuang murals, 2335 images of Gansu murals, 2950 images of Hebei murals, and 1482 images of Inner Mongolia murals, as illustrated in Fig. 4. All images are cropped to a resolution of 256 × 256 and divided into training, validation, and test sets in a ratio of 8:1:1.

We extract masks from authentic damage regions in the MaskCLP and MuralVerse datasets to simulate various damage patterns, such as cracks and blocky detachments. This approach introduces novel and realistic challenges to the field of image restoration.

The dataset MaskCLP-S, collected from publicly available online sources, comprises 8273 images of Chinese landscape paintings (as shown in Fig. 4), encompassing diverse artists and artistic styles. Within the dataset, 7446 images were used for training and 827 for testing, with all images uniformly preprocessed to a resolution of 256 × 256 pixels.

Implementation details

The experiments were conducted on a platform equipped with an NVIDIA GeForce RTX 3090 graphics card. Model training employed a stochastic gradient descent optimizer with an initial learning rate of 1×10^{-4} and a step size decay strategy. Considering the characteristics of different modules, the learning rate for the Global Structure Reasoning Module was set to 3×10^{-5}, the learning rate for the Semantic Stylistic Consistency Module was set to 5×10^{-4}, and the learning rate for the FGRM was set to 5×10^{-4}. The batch size is set to 8, and the total number of training epochs is 400.

During the training phase, all input images are uniformly resized to a resolution of 256 × 256. The hyperparameters in the loss function are set as follows: semantic loss weight $\lambda_1 = 0.5$, style loss weight $\lambda_2 = 0.2$. The ODE solution in the FGRM module employs the Euler method with a time step $\Delta t = 0.1$ and a total of 100 iterations.

Evaluation metrics

We follow the most common evaluation settings in image inpainting tasks, utilizing peak signal-to-noise ratio (PSNR), structural similarity index (SSIM)[26], and learned perceptual image patch similarity (LPIPS)[27] to assess the quality of image inpainting. Additionally, we employ image generation speed and the number of parameters as evaluation metrics for network performance, which intuitively reflect the model's complexity and execution speed.

Baselines

To demonstrate the effectiveness of our model, we selected seven representative baselines categorized into three groups.

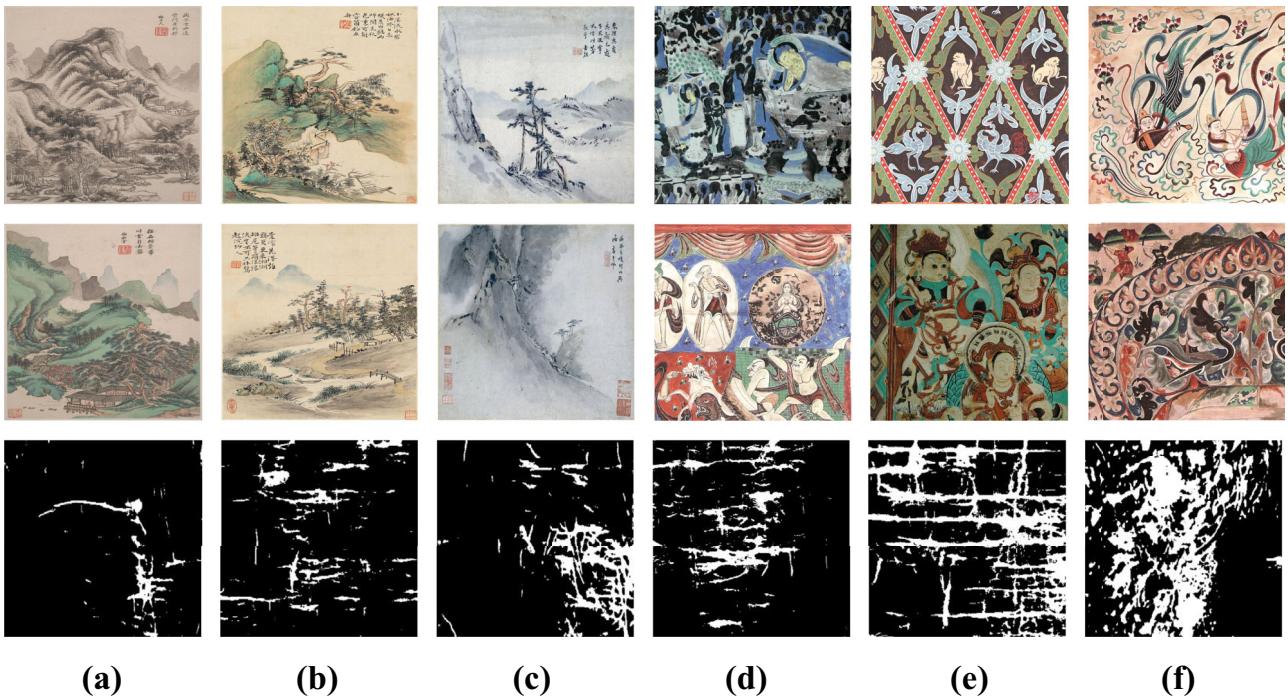


Fig. 4 | Illustration of the example of dataset and mask. a *Landscape Album with Clear Sounds* by *Jian Wang* from the early *Qing* dynasty, **b** *Twelve-Panel Landscape Album* by *Yan Hua* from the *Qing* dynasty, **c** *Landscape Album* by *Qipei Gao* from the *Qing* dynasty. **d**–**f** are famous Murals.

Table 1 | Model parameters and inference time

Model	Params	Δ*T*(ms)
CTSDG	52.5M	15.7
AdaIR	28.7M	68.3
EC	21.5M	7.1
RFR	31M	10.2
PromptIR	35.5M	32.6
Strdiffusion	42.7M	57665.8
RePaint	55.2M	13.6
Ours	54.1M	10

CNN-based inpainting.

• CTSDG[28]: A coupled texture-structure decomposition network implementing dual-stream inpainting through task-specific subnetworks.

• AdaIR[29]: An adaptive image inpainting network that handles diverse degradations through frequency-domain feature modulation and learnable degradation adaptation mechanisms.

• EdgeConnect[5]: A two-stage adversarial framework first reconstructing edge structures through semantic boundary detection, followed by texture completion guided by edge constraints.

• RFR[30]: A progressive inpainting framework employing iterative refinement with cascaded recurrent feedback modules.

Transformer-based inpainting.

• PromptIR[31]: A plug-and-play image inpainting framework enabling parameter-efficient integration with existing network architectures through prompt learning.

Diffusion-based inpainting.

• Strdiffusion[32]: A lightweight diffusion sampler with momentum-based skip for fast, multi-scale inpainting.

• RePaint[18]: A diffusion inpainting method that couples denoising sampling with mask-aware reverse SDE for structure and texture consistency.

Comparison analysis with SOTAs on MuralVerse-S

To demonstrate the effectiveness of the proposed model, AICE, we conducted qualitative and quantitative comparisons with several existing state-of-the-art methods.

Quantitative comparison with SOTAs on MuralVerse-S

For a rigorous evaluation of our proposed M3SFormer model, we compared it against various mainstream image inpainting models. For each compared model, we first fine-tuned it on our training set and then evaluated it on our test set. We did not impose specific constraints on the training hyperparameters for any model; instead, all models were trained using their original hyperparameters without modification. For the fine-tuning process, all models were trained for 10,000 steps. The parameter counts and inference times of the experimental models are shown in Table 1. For the test set, we categorized masks into three groups based on their area ratio in the image: 0.1%–10%, 10%–20%, and 20%–30%. For each image, one mask was randomly selected from each category and applied to the image. The final experimental results are presented in Table 2 and Fig. 5.

It can be observed that M3SFormer consistently outperforms competitors in PSNR, not only surpassing the formidable Repaint but also achieving a 31.677% improvement over PromptIR at 20–30% mask coverage. This demonstrates our model's ability to generate superior image quality, standing out notably among other benchmark models. M3SFormer also demonstrates exceptional performance on SSIM, achieving a 74.047% improvement over PromptIR[31], another Transformer-based architecture, at mask levels of 20–30%. This demonstrates that our model surpasses others in structural similarity. Simultaneously, our approach exhibits exceptional stability. For instance, the SSIM of the EC[5] model drops sharply as the mask coverage increases, with its metric at 20–30% decreasing by a full 25.080% compared to its metric at 0.1–10%. In contrast, our method metrics at 20–30% mask coverage decrease by only 0.521% compared to those at 0.1–10% coverage, demonstrating the exceptional stability of our model. In terms of parameter count and inference efficiency, our method achieves the second-fastest inference speed with a moderate model size. More importantly, it strikes an excellent balance between efficiency and generation

Table 2 | Comparison results of M3SFormer on MuralVerse-S

Model	Type	PSNR↑			SSIM↑			LPIPS↓		
		0.1–10%	10–20%	20–30%	0.1–10%	10–20%	20–30%	0.1–10%	10–20%	20–30%
CTSDG[28]	CNN	26.449	26.113	26.016	0.782	0.774	0.771	0.271	0.275	0.278
AdaIR[29]	CNN	23.585	22.562	21.268	0.917	0.875	0.853	0.082	0.119	0.141
EC[5]	CNN	25.907	20.183	18.455	0.933	0.741	0.699	0.047	0.201	0.231
RFR[30]	CNN	23.496	18.571	18.617	0.891	0.654	0.648	0.122	0.311	0.302
PromptIR[31]	Transformer	21.164	20.675	20.393	0.555	0.554	0.551	0.393	0.391	0.394
Strdiffusion[32]	Diffusion	25.826	25.622	25.021	0.898	0.881	0.873	0.042	0.042	0.045
RePaint[18]	Diffusion	27.474	26.981	26.433	0.931	0.901	0.893	0.026	0.029	**0.030**
Ours	Transformer	**27.891**	**27.304**	**26.853**	**0.959**	**0.956**	**0.954**	**0.015**	**0.022**	0.035

The output images of the generators are used for metrics computation. ↑ Higher values are better, ↓ Lower values are better. *Optimal results are displayed in **bold**, while suboptimal results are underlined.

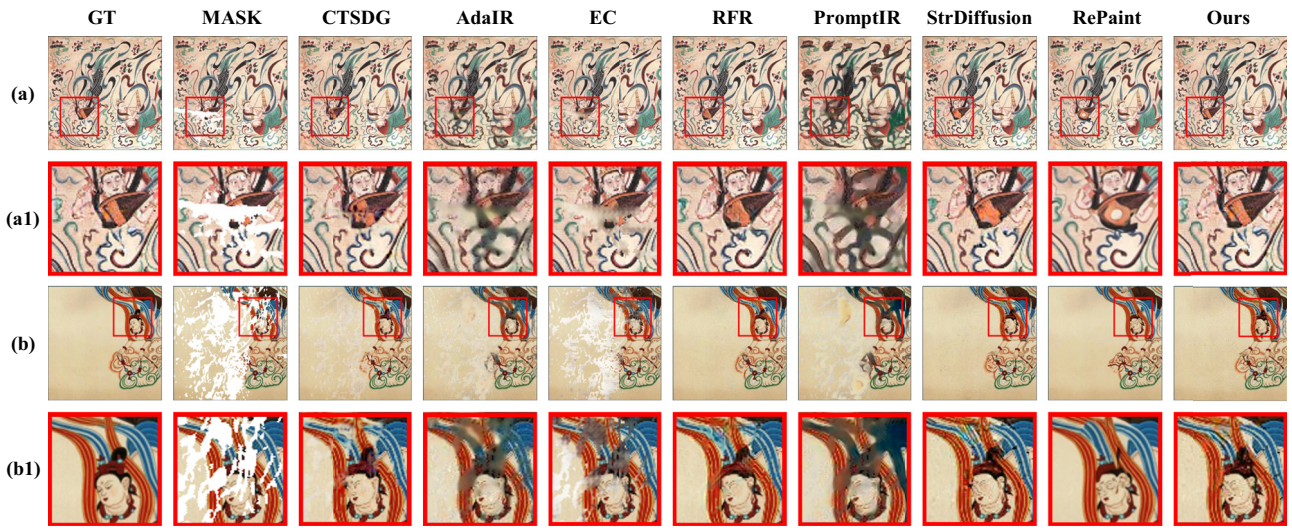


Fig. 5 | Illustration of traditional Chinese mural comparison. a, **b** Mural and results of the corresponding comparison method. **a1**, **b1** Enlarged view of the red area.

quality: it delivers superior output quality compared to the fastest model, EC, while requiring significantly fewer parameters and lower computational cost than the high-quality model RePaint.

M3SFormer also demonstrates outstanding performance on LPIPS, achieving a 91.116% improvement over PromptIR on masks ranging from 20 to 30%. This represents a significant advancement, indicating our model outperforms comparable architectures in enhancing perceptual similarity.

In summary, these experimental results demonstrate the validity and efficiency of our model, which excels across all metrics.

Qualitative analysis with SOTAs on MuralVerse-S

The comparative experimental results are shown in Fig. 6. Overall, the proposed model demonstrates comprehensively superior performance at the visual level. The inpainting results it generates are not only more structurally complete but also exhibit texture details that better align with the contextual semantics, thus more closely matching human visual perception habits and significantly outperforming other compared models in terms of overall esthetic effect.

Figure 6a1 shows an enlarged view of the red box region in Fig. 6a. It can be clearly observed that the results generated by EC[5] and PromptIR[31] suffer from overall blurring and structural distortion, particularly with broken edges, leading to poor inpainting quality. Although RFR[30] and StrDiffusion[32] manage to preserve the general structure, they exhibit unnatural clusters of black pixel artifacts in local areas, severely compromising the perceptual detail and authenticity. In contrast, the results generated by CTSDG[28], RePaint[18], and our method are visually closer to the original image. Among these, our model not only excels in detail completeness but also performs best in terms of overall visual consistency.

Figure 6b1 provides a locally magnified view of the red box region in Fig. 6b. The issues observed here are largely consistent with those mentioned above: the areas inpainted by EC and PromptIR[31] show noticeable color discrepancies compared to the surrounding original regions, indicating restoration inconsistency due to insufficient semantic understanding. RFR[30] and StrDiffusion[32] again exhibit issues with black pixel noise, further degrading the overall image quality. Our model effectively overcomes these typical drawbacks, with the inpainted regions demonstrating more natural and harmonious structural connections and color transitions with the surrounding content.

In summary, across diverse test scenarios, our model consistently exhibits comprehensive inpainting capabilities superior to other mainstream methods. Particularly in challenging areas where other models commonly suffer from blurring, structural breaks, noise interference, or color distortion, our model—leveraging its multi-stage fusion mechanism and style-aware design—still generates high-quality inpainting results that are structurally reasonable, stylistically consistent, and visually stable.

Quantitative comparison with SOTAs on MaskCLP-S

To evaluate the performance across different datasets, we conducted experiments on various types of datasets. For each dataset, we performed fine-tuning training and testing. The results are shown in Table 3.

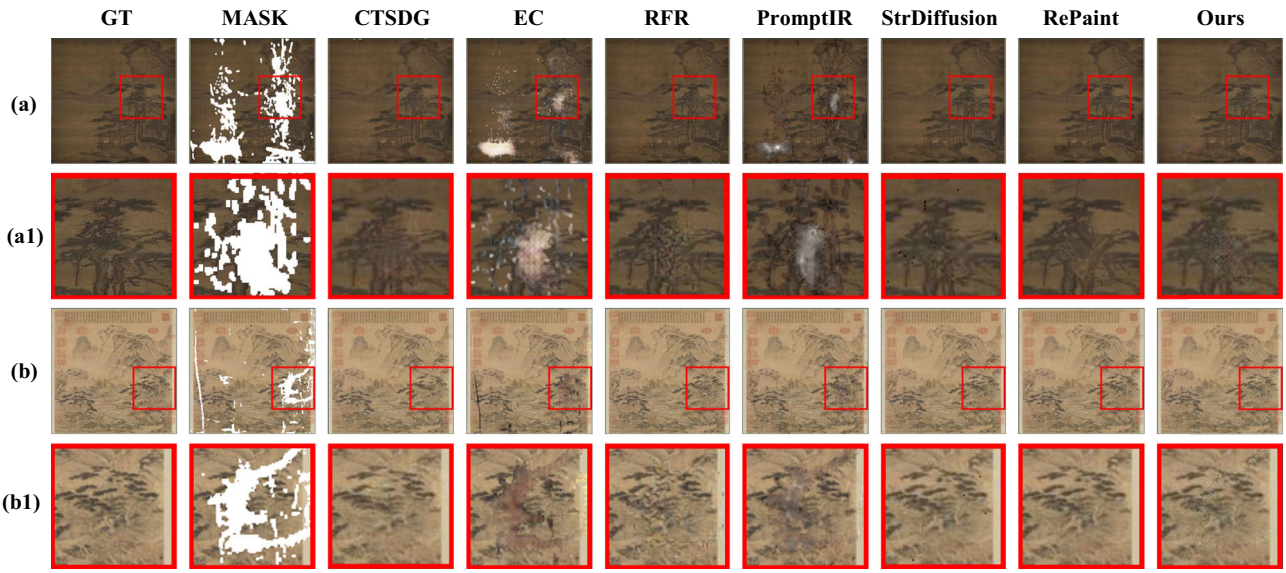


Fig. 6 | Illustration of traditional Chinese painting comparison. a, **b** Results of landscape analysis and related comparative methods, **a1**, **b1** Enlarged view of the red area.

Table 3 | Comparison results of M3SFormer on MaskCLP-S

Model	Type	PSNR↑			SSIM↑			LPIPS↓		
		0.1–10%	10–20%	20–30%	0.1–10%	10–20%	20–30%	0.1–10%	10–20%	20–30%
CTSDG[28]	CNN	26.479	26.146	26.216	0.782	0.774	0.771	0.271	0.275	0.278
EC[5]	CNN	25.927	20.283	18.655	0.933	0.741	0.699	0.047	0.201	0.231
RFR[30]	CNN	23.486	18.771	18.687	0.894	0.658	0.644	0.122	0.311	0.302
PromptIR[31]	Transformer	21.061	20.685	20.363	0.556	0.554	0.551	0.394	0.391	0.392
Strdiffusion[32]	Diffusion	25.814	25.601	24.936	0.898	0.882	0.871	0.042	0.043	0.045
RePaint[18]	Diffusion	27.452	26.884	26.285	0.932	0.900	0.892	0.025	0.029	**0.030**
Ours	Transformer	**27.863**	**27.266**	**26.762**	**0.958**	**0.956**	**0.955**	**0.017**	**0.023**	0.034

The output images of the generators are used for metrics computation. ↑ Higher values are better, ↓ Lower values are better. *Optimal results are displayed in **bold**, while suboptimal results are underlined.

It can be observed M3SFormer still leads in PSNR, outperforming its formidable competitor Repaint[18]. Moreover, it achieves a 31.424% improvement over PromptIR[31], another Transformer-based architecture, on masks ranging from 20% to 30%. This demonstrates that our model continues to generate superior image quality even when applied to a different dataset. M3SFormer also demonstrates exceptional performance on SSIM, achieving a 73.321% improvement over PromptIR[31], another Transformer-based architecture, at 20–30% mask coverage. This demonstrates that our model maintains superior structural similarity performance across diverse datasets. Additionally, our approach exhibits exceptional stability. At 20–30% mask coverage, our metrics decline by only 0.313% compared to those at 0.1–10% coverage, indicating outstanding consistency across various datasets.

M3SFormer also demonstrates outstanding performance on LPIPS, achieving a 91.326% improvement over PromptIR[31], another Transformer-based architecture, on masks ranging from 20 to 30%. This represents a significant advancement, indicating our model outperforms comparable architectures in enhancing perceptual similarity.

In summary, these experimental results demonstrate the generalizability of our model.

Qualitative analysis with SOTAs on MaskCLP-S

The comparison results are shown in Fig. 6. It can be observed that the proposed model demonstrates superior performance at the visual level, with generated inpainting results more aligned with human perception and significantly outperforming other comparison models in overall quality.

Figure 6a1 presents an enlarged view. It is evident that the results generated by EC[5] and PromptIR[31] exhibit noticeable blurring and structural distortion, resulting in poor inpainting quality. RFR[30] and StrDiffusion[32], meanwhile, show localized clusters of black pixels that compromise the perception of fine details. In contrast, the results produced by CTSDG[28], RePaint[18], and our proposed method are visually closer to the original image, with our model demonstrating the best performance in terms of detail preservation and visual consistency.

Figure 6b1 shows a local enlargement of the red-boxed area. The results are similar to the previous case: the colors in the restored regions of EC[5] and PromptIR[31] show noticeable discrepancies with the surrounding areas, resulting in inconsistent inpainting. RFR[30] and StrDiffusion[32] also exhibit black pixel noise issues, affecting overall image quality. The proposed model effectively avoids these issues, achieving natural transitions in both structure and color between the restored region and its surroundings.

In summary, across all test scenarios, the proposed model demonstrates superior inpainting capabilities compared to other methods. Notably, even when other models produce blurred, structurally broken, or color-distorted results, the proposed model consistently generates structurally coherent, stylistically consistent, and visually stable inpaintings.

Visualization on celebrated traditional mural painting datasets

To validate the effectiveness in practical scenarios, we cropped images from the *Fahai Temple murals* and manually added random masks as inpainting targets. The inpainting results are shown in Fig. 7. The experiment demonstrates that even under random mask interference, the proposed

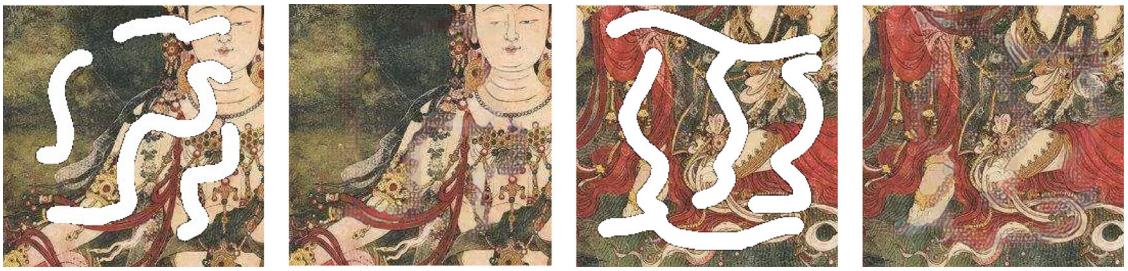


Fig. 7 | Illustration of celebrated traditional mural painting. a, **b** Masking a section of the *Fahai Temple murals*. **a1**, **b1** Inpainting result of a section of the *Fahai Temple murals*.

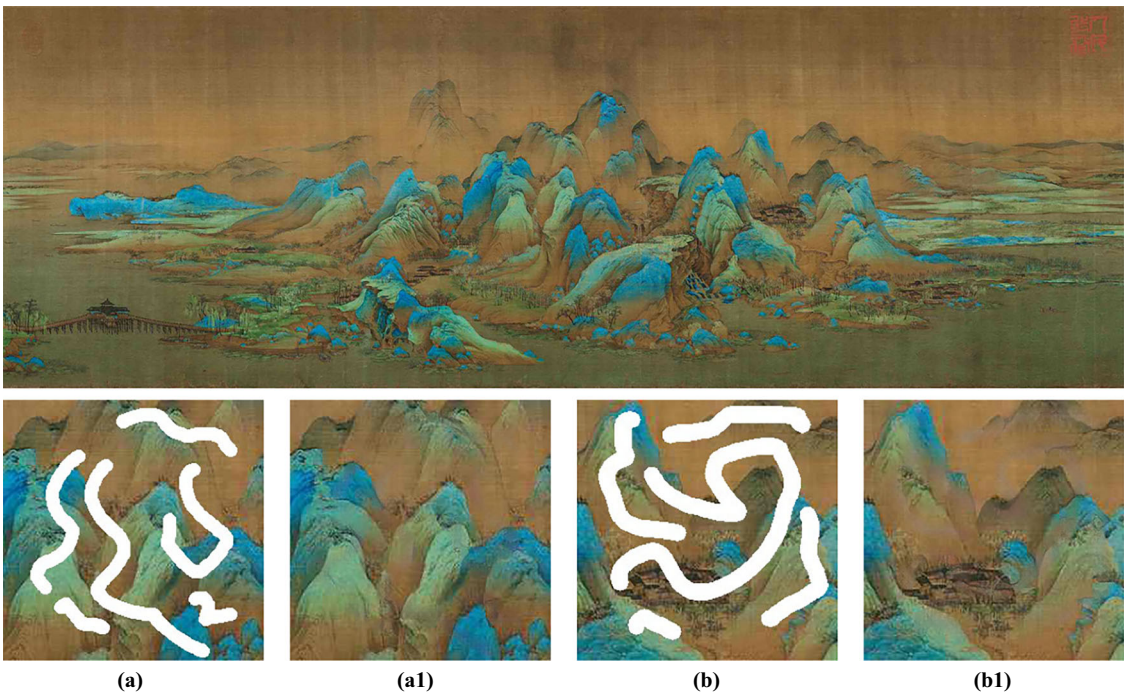


Fig. 8 | Illustration of traditional celebrated Chinese painting. a, **b** Applying a mask to a section of *the Thousand Miles of Rivers and Mountains scroll*. **a1**, **b1** The inpainting Result of the section of *the Thousand Miles of Rivers and Mountains scroll*.

model can effectively complete mural inpainting tasks. Although minor color differences between the restored area and its surroundings are visible upon magnification, the model demonstrates strong inpainting capabilities in preserving overall content structure and color harmony.

These results indicate that our method effectively understands image semantic content and can reasonably infer the structure and color of missing areas based on contextual information, showing potential for tackling such complex inpainting tasks.

Visualization on celebrated traditional chinese painting datasets

To validate the generalization capability across diverse artistic styles, we further selected the *Thousand Miles of Rivers and Mountains scroll* as a test subject. After cropping the image and manually adding random masks, we applied our proposed method for inpainting. The results are shown in Fig. 8.

Experiments demonstrate that even under random mask interference, the model can effectively recover the content of the ancient painting. Although minor color differences between the restored area and its surroundings are visible upon magnification, the overall structural reconstruction is accurate. Colors blend naturally with the surrounding environment, with no noticeable visual discontinuities. This result further demonstrates that our model can fully comprehend the rich semantic information within images. By leveraging contextual features, it effectively infers the reasonable structure and color distribution of missing regions, showcasing excellent generalization and practicality.

Ablation study

To validate the contributions of each component within our framework, we conducted ablation experiments across the following four dimensions.

Ablation study on effectiveness of components

This experiment systematically evaluated the contribution of each module to model performance by sequentially removing the FGRM, GSRF, and SSCM. The results are shown in Table 4.

The results indicate that removing any module leads to a decline in image inpainting quality. Specifically, the PSNR value decreases significantly as the masked area expands. At a masking ratio of 20–30%, removing any module causes a PSNR drop ranging from 1% to 2.5%, with the removal of GSRF having a relatively smaller impact on PSNR. Regarding structural similarity, SSIM slightly decreased at smaller mask ratios 0.1–10%, but its performance gradually improved as the mask ratio increased.

All three modules contributed approximately a 0.15% improvement in SSIM under large mask conditions, indicating that each module helps

Table 4 | Ablation of structural components and training strategies

Mask	Method	PSNR↑	SSIM↑	LPIPS↓
0.1–10%	w/o GSRF	27.586	0.961	0.014
	w/o FGRM	27.580	**0.962**	0.014
	w/o SSCM	27.538	0.960	**0.011**
	Ours	**27.891**	0.959	0.015
10–20%	w/o GSRF	27.141	0.957	**0.019**
	w/o FGRM	26.760	0.954	0.020
	w/o SSCM	26.824	**0.959**	0.019
	Ours	**27.304**	0.956	0.022
20–30%	w/o GSRF	26.578	0.952	**0.026**
	w/o FGRM	26.191	0.950	0.028
	w/o SSCM	26.184	0.953	0.031
	Ours	**26.853**	**0.954**	0.035

Optimal results are displayed in bold, while suboptimal results are underlined.

enhance the structural consistency between the restored result and the original image.

Figure 9c1–f1 further reveals that removing any module induces local generation artifacts. As shown in the third row of Fig. 9c1–e1, an unreasonable brown patch appears slightly left of the image center.

In summary, all three module play crucial roles in enhancing the model inpainting performance. GSRF strengthens the ability to capture long-range dependencies in images, effectively preserving high-frequency details. SSCM provides dual constraints at both semantic and stylistic levels during inpainting. FGRM further improves visual coherence and structural integrity through flow-guided fine-grained optimization.

Ablation Study on effectiveness of feature representation

This experiment systematically compares the impact of quantitative features (QF) and continuous features (CF) on model performance, validating the effectiveness of the improved module. Relevant results are shown in Table 5.

The experimental results reveal that using the original quantized features yields only a marginal improvement in PSNR, with SSIM remaining largely unchanged. Conversely, LPIPS shows a significant decline under larger mask ratios. Specifically, at mask ratios of 10%-20%, LPIPS improves by 37.142%. A substantial 43.548% improvement at 20%-30% mask coverage.

Thus, continuous features outperform quantized features in improving perceptual quality. prioritizing image quality metrics alone leads to a significant reduction in perceptual similarity.

Further visual comparisons in Fig. 10c1, d1 reveals that images generated using quantization features, as shown in Fig. 10c1, exhibit white cracks, compromising visual naturalness. In contrast, as shown in Fig. 10d1, results using continuous features exhibit reasonable filling of these cracks. Although colors are not perfectly matched with surrounding areas, the overall appearance is softer and more natural, aligning better with human visual perception habits.

In summary, incorporating continuous features enables the model to better learn deep characteristics within images, resulting in inpainting outcomes that appear more visually natural and gentle. This approach aligns more closely with human expectations for perceiving image content.

Ablation study on effectiveness of step size

This experiment systematically evaluated the impact of different ODE solution step sizes on model performance. The results are summarized in Table 6.

Experimental results indicate that as the number of ODE steps increases, all three metrics—PSNR, SSIM, and LPIPS, show varying degrees of improvement. The most significant enhancement in PSNR occurs when the number of steps reaches 50. For mask ratios of 0.1–10%, 10–20%, and 20–30%, the PSNR gains from 20 to 50 steps were 4.063%, 2.465%, and 1.439%, respectively. This demonstrates that appropriately increasing the ODE step count comprehensively enhances model performance. Based on this, setting the ODE step count to 50 achieves favorable results.

Figure 11c1–e1 further illustrates that image generation quality progressively improves with increasing ODE steps. For instance, in Fig. 11c1, color blocks appear in the facial region that clash with the surrounding hues, whereas in Fig. 11e1, these blocks are noticeably softened, resulting in a more harmonious and natural overall image Fig. 12.

In summary, appropriately increasing the ODE step size positively impacts model performance.

Ablation study on effectiveness of loss function

This experiment systematically evaluated the impact of each loss function on model performance by sequentially removing the hierarchical style loss, semantic segmentation loss, and both losses simultaneously. The results are shown in Table 7.

The results indicate that removing any single loss function leads to a decrease in PSNR, with the most significant impact observed when both style loss and semantic segmentation loss are removed simultaneously. Under three mask ratios, 0.1–10%, 10–20%, and 20–30%, PSNR decreased by 8.941%, 7.053%, and 4.368%, respectively. Regarding LPIPS, while removing the style loss yields some improvement, the other two settings result in decreased perceptual similarity. For structural similarity, SSIM decreases after removing any loss function, indicating that all three are crucial for maintaining image structural consistency.

Further observation of Fig. 11c1–f1 reveals that removing any loss function degrades generation quality. As shown in the top row of Fig. 11d1 and Fig. 11e1, white artifacts appear in the upper part of the image, inconsistent with the surrounding areas. While retaining only partial losses, as shown in Fig. 11c1, mitigates the white artifact issue, local brightness inconsistencies persist in the facial region. Only when all loss functions are fully applied, as shown in Fig. 11f1, yielding generation results that are more reasonable in both structure and detail.

In summary, style loss, semantic segmentation loss, and their combined effect significantly impact model performance. Their synergistic interaction effectively enhances the quality and structural consistency of generated images. Although incorporating these losses slightly affects perceptual similarity metrics, sacrificing them to pursue higher LPIPS scores would result in a significant decline in overall visual quality. Therefore, such a trade-off is not advisable in the overall evaluation.

Discussion on failure cases

As illustrated in Fig. 13, the failure cases indicate that our method does not perform well in repairing relatively large areas or regions with complex structures. For example, Fig. 13a exhibits noticeable color patches that are inconsistent with the surrounding context, while in Fig. 13c, an object inconsistent with the context is generated in the mouth area of the figure. We believe these issues can be largely attributed to the model's hallucination problem when processing highly uncertain regions. When sufficient surrounding pixel constraints are absent, the model tends to rely on common features from its training data for completion, which can easily lead to factually divergent results in complex scenarios requiring high detail. A core issue lies in the failure of the current model to leverage prompts as a powerful guiding mechanism. By introducing prompt inputs—whether in the form of textual descriptions, edge maps, or even rough user feedback—the highly uncertain "open-loop" generation problem can be transformed into a "closed-loop" optimization problem guided by strong priors. Such guidance can effectively anchor the model's generation space, mitigating hallucinations at the source and ensuring structural consistency with the overall image as well as historical credibility of the restored content. In terms of specific strategy, prompts can be designed to influence internal modules of the AQ-Transformer, so that the data ultimately fed into the SSCM carries prompt information, thereby enabling controlled regulation of the final output.

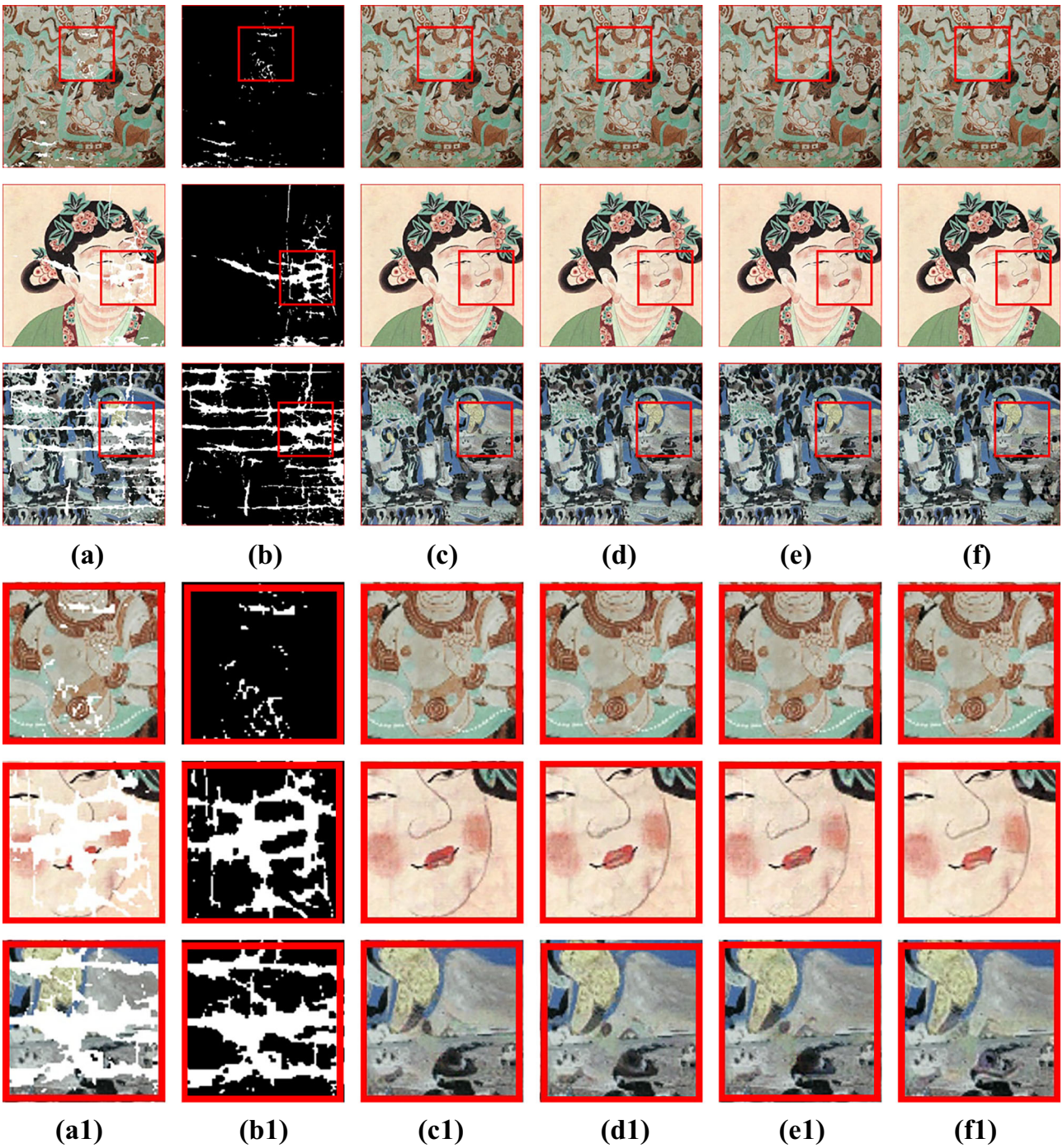


Fig. 9 | Illustration of the Components Ablation Study. a 0.1–10%, 10–20%, 20–30% masks combined with the mural. **b** Mask. **c** Removing GSRF. **d** Removing FGRM. **e** Removing SSCM. **f** Ours. **a1**–**f1** Enlarged views of the red regions.

Table 5 | Ablation of feature representation strategies

Mask	Method	PSNR↑	SSIM↑	LPIPS↓
0.1–10%	w/ QF	27.403	0.955	0.021
	w/CF	**27.891**	**0.959**	**0.015**
10–20%	w/ QF	27.075	0.954	0.035
	w/CF	**27.304**	**0.956**	**0.022**
20–30%	w/QF	**26.856**	**0.953**	0.062
	w/CF	26.853	**0.954**	**0.035**

QF denotes quantized features, while CF denotes continuous features.
Optimal results are displayed in bold, while suboptimal results are underlined.

Discussion

Our approach does not perform well in repairing large-scale missing areas or inherently complex regions, often yielding unreasonable results. These issues stem from multiple factors, including the absence of prompt-like inputs in our model. Without such prompt-guided constraints, it becomes challenging to generate visually coherent outputs in certain scenarios.

This study not only provides a scalable technical approach for the digital inpainting of ancient murals but also offers new insights for the field of digital preservation of cultural relics. Future work will focus on further optimizing model efficiency, exploring adaptive inpainting mechanisms for cross-cultural mural styles, and extending this technical framework to more types of cultural heritage preservation scenarios.

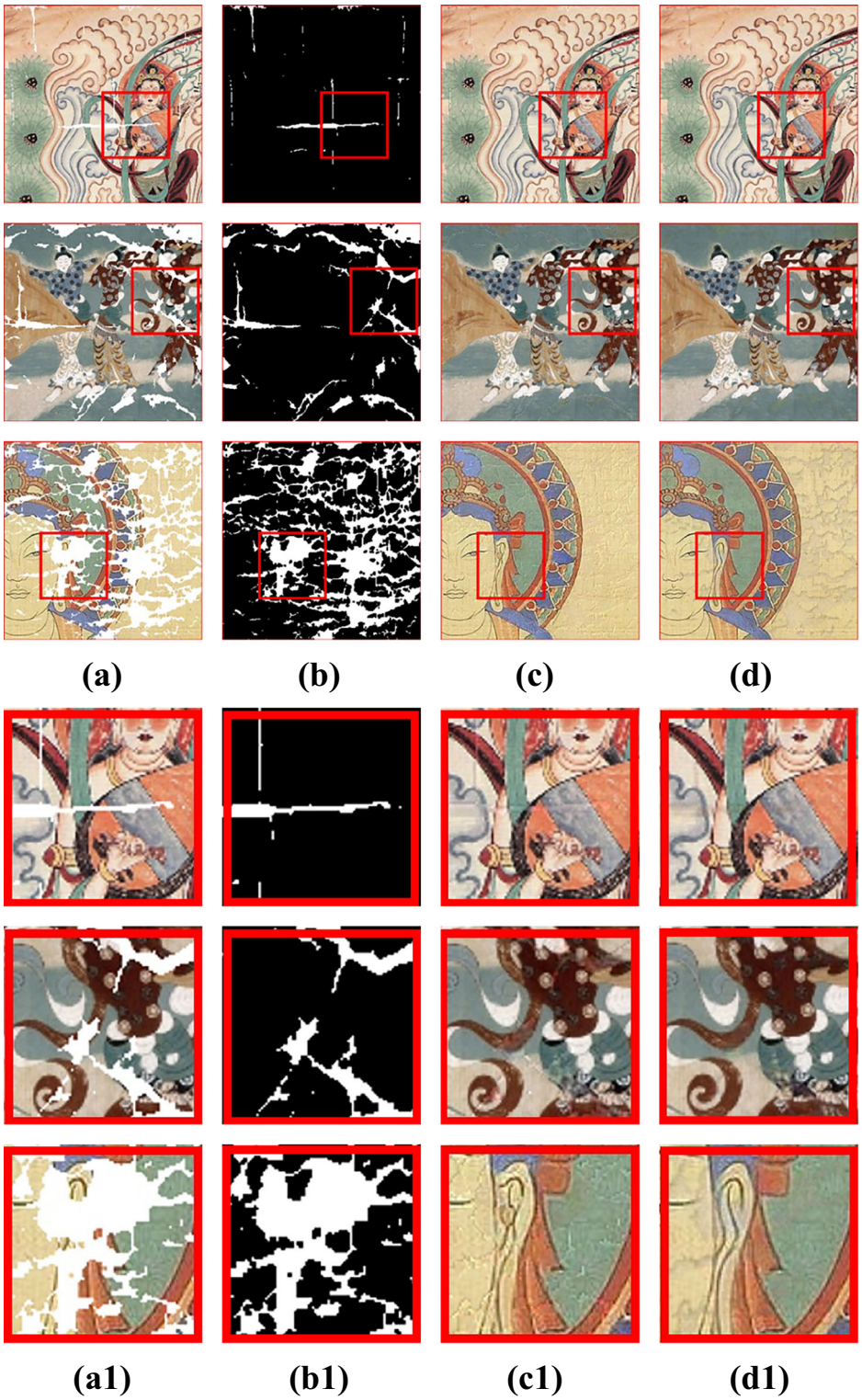


Fig. 10 | Illustration of feature representation ablation study. a Images combining masks with mural paintings at 0.1–10%, 10–20%, and 20–30% coverage. **b** Mask. **c** w/QF. **d** w/CF. **a1**–**d1** Enlarged views, respectively.

This study addresses key challenges in the digital inpainting of ancient murals by proposing M3SFormer, a mural image inpainting framework integrating multi-stage optimization strategies with a semantic-stylistic consistency guidance mechanism. By establishing a multi-stage inpainting workflow progressing from coarse to fine-grained processing and introducing a guidance mechanism based on refined flow field prediction, it effectively resolves issues of semantic misalignment and structural discontinuity encountered by traditional methods when handling large-scale missing areas. Experimental results demonstrate that M3SFormer exhibits significant advantages across multiple mural image datasets, particularly in restoring intricate details within complex regions and preserving artistic style, achieving notable improvements over existing mainstream methods.

The core contribution of this study lies in the innovative introduction of continuous feature modeling strategies to mural restoration, abandoning traditional discrete quantization methods. By adopting the UQ-Transformer component of the improved P-VQVAE, this approach preserves richer detail features and texture information, significantly enhancing the restoration quality of high-frequency structures. Simultaneously, the

proposed SSCM model effectively resolves color and artistic style inconsistencies between filled regions and the original image by integrating regional semantic information with multi-layer perceptual style features. Furthermore, the FGRM mechanism enables fine-grained adjustments to the filling results, ensuring structural consistency and visual coherence

Table 6 | Effectiveness of ODE step sizes in FGRM

Mask Ratio	T-step	PSNR↑	SSIM↑	LPIPS↓
0.1–10%	20	26.802	**0.960**	**0.014**
	50	**27.891**	0.959	0.015
10–20%	20	26.647	**0.957**	0.023
	50	**27.304**	0.956	**0.022**
20–30%	20	26.472	0.953	**0.034**
	50	**26.853**	**0.954**	0.035

T-Step is 20, 50, 100, respectively
Optimal results are displayed in bold, while suboptimal results are underlined.

Table 7 | Effectiveness of loss components

Mask ratio	Method	PSNR↑	SSIM↑	LPIPS↓
0.1–10%	w/o Style Loss	27.705	**0.962**	0.015
	w/o Sem Loss	26.403	0.955	**0.014**
	w/o Both Loss	25.397	0.952	0.016
	Ours	**27.891**	0.959	0.015
10–20%	w/o Style Loss	27.226	**0.957**	**0.020**
	w/o Sem Loss	26.052	0.951	0.025
	w/o Both Loss	25.378	0.953	0.025
	Ours	**27.304**	0.956	0.022
20–30%	w/o Style Loss	26.607	0.952	**0.027**
	w/o Sem Loss	26.319	0.951	0.043
	w/o Both Loss	25.680	0.951	0.040
	Ours	**26.853**	**0.954**	0.035

Optimal results are displayed in bold, while suboptimal results are underlined.

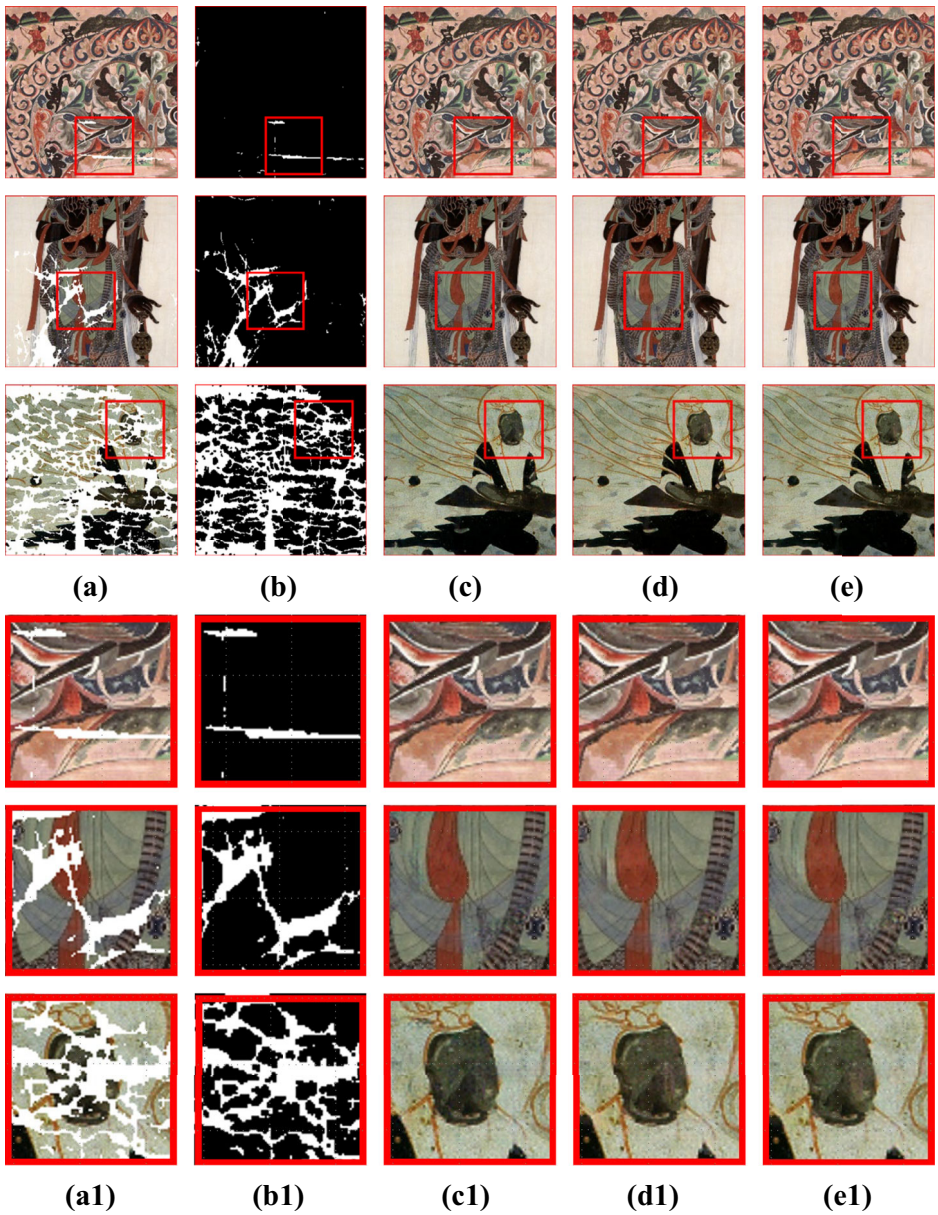


Fig. 11 | Illustration of step size ablation study. **a** Images combining masks with wall paintings at 0.1–10%, 10–20%, and 20–30% coverage. **b** Mask. **c** ODE step = 20. **d** ODE step = 50. **e** ODE step = 100.

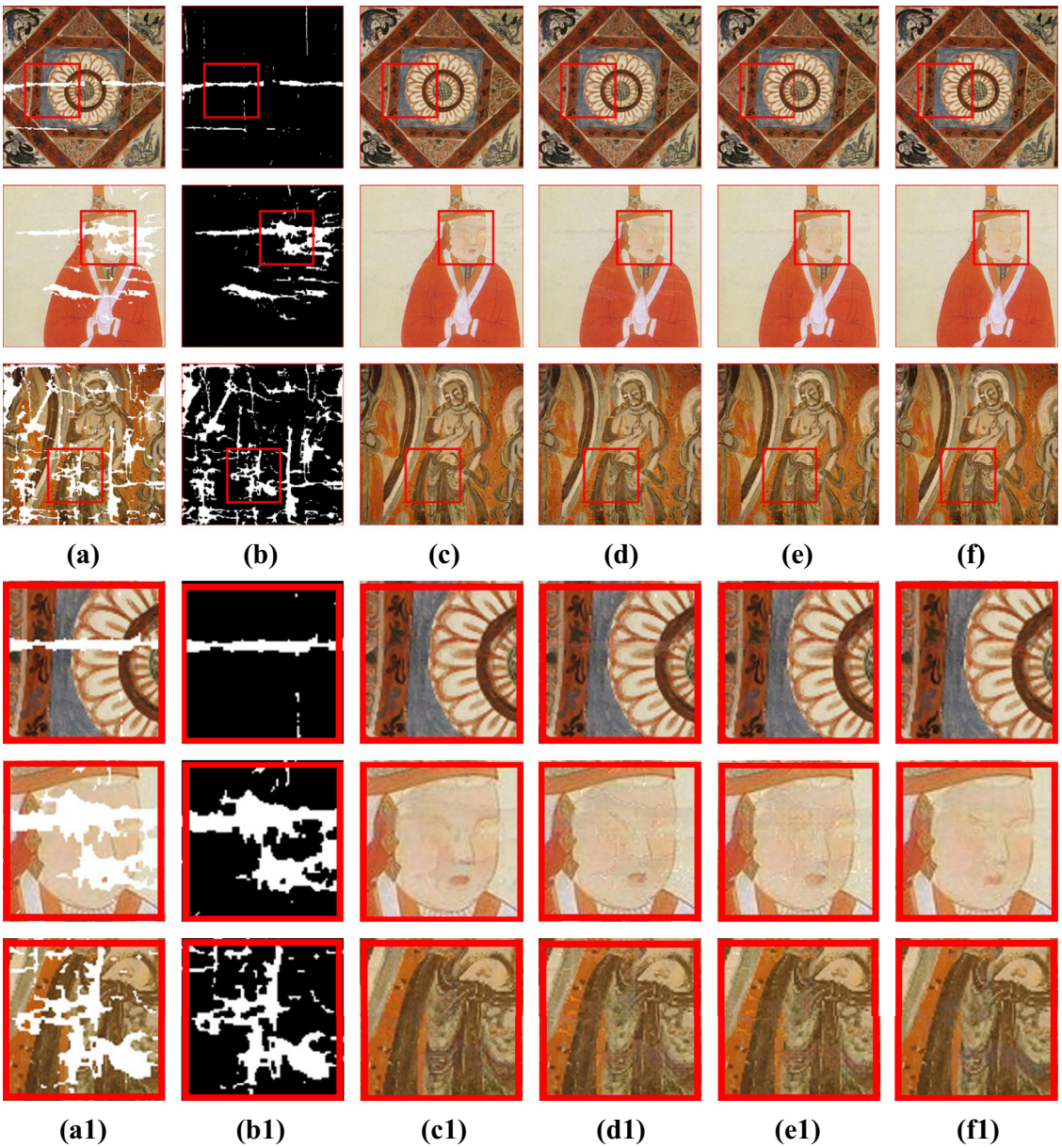


Fig. 12 | Illustration of loss function ablation study. a 0.1–10%, 10–20%, 20–30% masks combined with murals. **b** Mask. **c** Removing style loss. **d** Removing sem loss. **e** Removing both loss. **f** Ours. **a1**–**f1** Enlarged views of the red region.

Fig. 13 | Illustration of failure cases. a, **b** Show the mural before inpainting, while **a1**, **b1** illustrate examples of inpainting errors.

throughout the restoration process. Research findings indicate that by deeply integrating deep learning technology with the requirements of cultural heritage preservation, inpainting quality can be effectively enhanced, creating new possibilities for the permanent preservation and dissemination of cultural heritage.

Data availability

The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request. The dataset in this study is available at https://github.com/LPDLG/M3SFormer.

Code availability

The code used in this study is available from the corresponding author upon reasonable request.

Received: 29 September 2025; Accepted: 13 January 2026;
Published online: 28 January 2026

References

1. Li, J., Wang, N., Zhang, L., Du, B. & Tao, D. Recurrent feature reasoning for image inpainting. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7760–7768 (IEEE, 2020).
2. Li, Y., Zhang, C., Li, Y., Sui, D. & Guo, M. An improved mural image restoration method based on diffusion model. *npj Herit. Sci.* **13**, 347 (2025).
3. Guan, J. et al. Progressive generative mural image restoration based on adversarial structure learning. *npj Herit. Sci.* **13**, 309 (2025).
4. Suvorov, R. et al. Resolution-robust large mask inpainting with Fourier convolutions. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, 2149–2159 (IEEE, 2022).
5. Nazeri, K., Ng, E., Joseph, T., Qureshi, F. Z. & Ebrahimi, M. Edgeconnect: generative image inpainting with adversarial edge learning. Preprint at https://doi.org/10.48550/arXiv.1901.00212 (2020).
6. Shen, J., Liu, N., Sun, H., Li, D. & Zhang, Y. An instrument indication acquisition algorithm based on lightweight deep convolutional neural network and hybrid attention fine-grained features. *IEEE Trans. Instrum. Meas.* **73**, 1–16 (2024).
7. Shen, J. et al. Finger vein recognition algorithm based on lightweight deep convolutional neural network. *IEEE Trans. Instrum. Meas.* **71**, 1–13 (2022).
8. Shen, J. et al. An anchor-free lightweight deep convolutional network for vehicle detection in aerial images. *IEEE Trans. Intell. Transp. Syst.* **23**, 24330–24342 (2022).
9. Liang, J. et al. Clusterfomer: clustering as a universal visual learner. *Adv. Neural Inf. Process. Syst.* **36**, 64029–64042 (2023).
10. Liang, J., Zhou, T., Liu, D. & Wang, W. Clustseg: clustering for universal segmentation. Preprint at https://doi.org/10.48550/arXiv.2305.02187 (2023).
11. Liu, D. et al. Densernet: weakly supervised visual localization using multi-scale feature aggregation. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 35, 6101–6109 (AAAI Press, 2021).
12. Shen, J. et al. An algorithm based on lightweight semantic features for ancient mural element object detection. *npj Herit. Sci.* **13**, 70 (2025).
13. Lingle, L. D. Transformer-vq: linear-time transformers via vector quantization. Preprint at https://doi.org/10.48550/arXiv.2309.16354 (2023).
14. Barnes, C., Shechtman, E., Finkelstein, A. & Goldman, D. B. Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* **28**, 24 (2009).
15. Liu, Q. et al. Transformer based pluralistic image completion with reduced information loss. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 6652–6668 (2024).
16. Liao, L., Xiao, J., Wang, Z., Lin, C.-W. & Satoh, S. Guidance and evaluation: Semantic-aware image inpainting for mixed scenes. In *Proc. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, 683–700 (Springer, 2020).
17. Li, L. et al. Line drawing guided progressive inpainting of mural damages. Preprint at https://doi.org/10.48550/arXiv.2211.06649 (2022).
18. Lugmayr, A. et al. Repaint: inpainting using denoising diffusion probabilistic models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 11461–11471 (IEEE, 2022).
19. Liu, X., Gong, C. & Liu, Q. Flow straight and fast: learning to generate and transfer data with rectified flow. Preprint at https://doi.org/10.48550/arXiv.2209.03003 (2022).
20. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE International Conference on Computer Vision* 2223–2232 (IEEE, 2017).
21. Deng, X. & Yu, Y. Ancient mural inpainting via structure information guided two-branch model. *npj Herit. Sci.* **11**, 131 (2023).
22. Chi, L., Jiang, B. & Mu, Y. Fast fourier convolution. *Adv. Neural Inf. Process. Syst.* **33**, 4479–4488 (2020).
23. Cheng, B., Misra, I., Schwing, A. G., Kirillov, A. & Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 1290–1299 (IEEE, 2022).
24. Kwatra, V., Essa, I., Bobick, A. & Kwatra, N. Texture optimization for example-based synthesis. In *ACM Siggraph 2005 Papers*, 795–802 (ACM, 2005).
25. Darabi, S., Shechtman, E., Barnes, C., Goldman, D. B. & Sen, P. Image melding: combining inconsistent images using patch-based synthesis. *ACM Trans. Graph.* **31**, 1–10 (2012).
26. Wang, Z., Bovik, A., Sheikh, H. & Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
27. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 586–595 (IEEE, 2018).
28. Guo, X., Yang, H. & Huang, D. Image inpainting via conditional texture and structure dual generation. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* 14134–14143 (IEEE, 2021).
29. Cui, Y. et al. AdaIR: adaptive all-in-one image restoration via frequency mining and modulation. In *Proc. The Thirteenth International Conference on Learning Representations* (ICLR 2025).
30. Li, J., Wang, N., Zhang, L., Du, B. & Tao, D. Recurrent feature reasoning for image inpainting. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2020).
31. Potlapalli, V., Zamir, S. W., Khan, S. & Khan, F. S. Promptir: prompting for all-in-one blind image restoration. Preprint at https://doi.org/10.48550/arXiv.2306.13090 (2023).
32. Liu, H., Wang, Y., Qian, B., Wang, M. & Rui, Y. Structure matters: tackling the semantic discrepancy in diffusion models for image inpainting. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 8038–8047 (IEEE, 2024).

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Nos. 62471390, 62406247, and 62306237), Key Project of Scientific Research Plan of Shaanxi Provincial Department of Education (No. 24JS052), and Key Laboratory of Archaeological Exploration and Cultural Heritage Conservation Technology (Northwestern Polytechnical University, No. 2024KFT03).

Author contributions

Q.Y. Hu: Conceptualization, software, validation, resources, data curation, Formal analysis. Q.F. Ge: Preparation, methodology, writing. Y.H. Zhang: Preparation, methodology. X.L. Peng: Software, investigation, validation. J.P. Wang: Administration. S.Y. Qu: Editing, supervision, project administration. N.N. Chen: Project administration. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Xianlin Peng or Nana Chen.