npj | heritage science

Article

https://doi.org/10.1038/s40494-026-02327-8

DCADif: decoupled conditional adaptive time-dynamic fusion diffusion inpainting of traditional Chinese mural paintings

Check for updates

Xianlin Peng[1,2], Chao Li[3], Qiyao Hu[3,4] ✉, Zengguo Sun[5], Jinye Peng[3,6] & Manli Sun[7] ✉

Digital inpainting of traditional Chinese murals is challenged by the difficulty of disentangling intricate structures from unique artistic styles, often leading to artifacts. To address this, we propose DCADif, a novel diffusion model for high-fidelity mural restoration. DCADif's core innovation is a Decoupled Conditional Encoder that uses parallel pathways a pre-trained CLIP for structural line art and a new SwinStyle Encoder for stylistic features to achieve independent control. Furthermore, a Time-Adaptive Feature Fusion (TAFF) module dynamically adjusts the influence of these features during denoising, prioritizing structure in early stages and style in later ones, mimicking an expert's coarse-to-fine workflow. Evaluated on our new large-scale MuralVerse-S dataset, DCADif significantly outperforms state-of-the-art methods across all degradation levels. It establishes a new benchmark for digital cultural heritage preservation by effectively balancing structural accuracy with artistic authenticity. The dataset and code are publicly available.The dataset and code are available at https://github.com/LPDLG/DCADif.

Traditional Chinese murals, as vital carriers of Chinese civilization, hold profound cultural significance. These murals, typically found on the walls of temples, grottoes, and tombs, document the religious beliefs, lifestyles, and aesthetic pursuits of ancient societies. They also embody the spiritual heritage of the Chinese nation, a legacy spanning millennia. However, these invaluable artworks are susceptible to damage and deterioration due to natural erosion and the passage of time. This situation underscores the critical importance of their preservation and inpainting. Figure 1 is the artwork *Mural from Fengguo Temple*, painted by anonymous artists of the *Yuan* Dynasty. The mural's unique style and texture render manual inpainting exceptionally challenging.

Deep learning-based image inpainting techniques now constitute a primary approach within computer vision for recovering missing information. They exhibit considerable promise in the digital preservation of cultural heritage[1]. Unlike conventional physical inpainting, this non-invasive approach enables high-fidelity content generation and texture reconstruction in damaged regions while preserving the integrity of the original artifact. This characteristic is particularly valuable for traditional Chinese murals, given their structural complexity, material fragility, and irreplaceability. Moreover, this technological domain is continuously advancing. In a seminal study, Yu et al. introduced the gated convolution method[2]. They observed that standard convolutional networks fail to differentiate between valid pixels and invalid values in damaged regions. This deficiency often results in inpainting marred by color distortion and structural artifacts. To overcome this limitation, they designed a learnable gating mechanism that allows the network to selectively process features from valid areas while disregarding corrupted ones. This innovation significantly enhanced the model's ability to handle large, irregularly shaped defects, producing results with more coherent structures and smoother textural transitions.

The research focus in both academia and industry has recently shifted towards a new paradigm: diffusion models. Prominent examples, such as Latent Diffusion Models[3] by Rombach et al. and the RePaint[4] by Lugmayr et al., have established a novel generative pathway. The core principle of this approach is an iterative denoising process that transforms random noise into high-quality content coherent with the surrounding image context. This method yields content with exceptional photorealism and detail. It also overcomes critical issues of training instability and mode collapse, which are inherent in Generative Adversarial Networks. These advancements establish diffusion models as the current state of the art in image generation and inpainting.

[1]School of Art, Northwest University, Xi'an, China. [2]State-Province Joint Engineering and Research Center of Advanced Networking and Intelligent Information Services, Xi'an, China. [3]School of Electronic Information, Northwest University, Xi'an, China. [4]Shaanxi Silk Road Cultural Heritage Digital Protection and Inheritance Collaborative Innovation Center, Xi'an, China. [5]School of Artificial Intelligence and Computer Science, Shaanxi Normal University, Xi'an, China. [6]Shaanxi Key Laboratory of Higher Education Institution of Generative Artificial Intelligence and Mixed Reality, Xi'an, China. [7]School of Culture Heritage, Northwest University, Xi'an, China. ✉e-mail: huqiyao@nwu.edu.cn; sunml68@sohu.com

Fig. 1 | Examples of *Mural from Fengguo Temple*, painted by anonymous artists of the *Yuan* Dynasty. These artworks are characterized by their complex compositions, distinctive artistic style, and significant physical degradation.

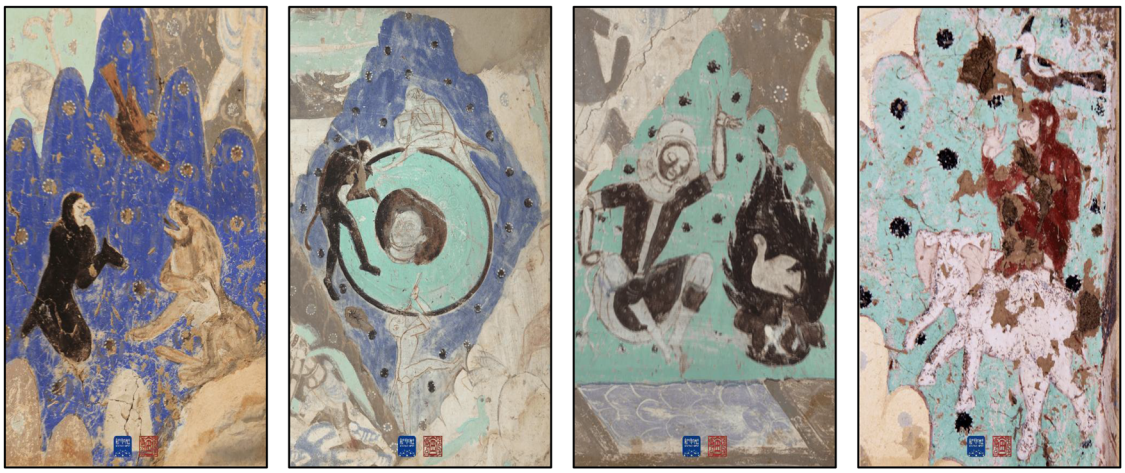

Fig. 2 | The Mural is the artwork *Mural from the Kizil Grottoes*, from the *Kucha Kingdom* period, created by anonymous artists. Characterized by their unique use of color, abstract motifs, and distinct shading techniques, these murals represent an aesthetic system different from traditional Central Plains styles.

However, the inpainting of traditional Chinese murals presents unique difficulties, as they are often characterized by complex compositions and abstract meanings. Figure 2 illustrates this with a mural from the Kizil Grottoes in Xinjiang. The artistic style of these works is defined by an intricate interplay of color, brushwork, and material texture. Thus, the primary challenge is to faithfully restore original details while preserving the distinctive artistic style. The key challenges in Chinese mural inpainting therefore encompass the following:

A primary challenge in image inpainting is the tendency to conflate structural reconstruction with stylistic information from a reference image[5]. This entanglement impedes the simultaneous achievement of structural accuracy and style fidelity, thereby compromising the inpainting overall quality and controllability.

The artistic style of traditional Chinese murals emerges from a complex interplay of color, brushwork, and material texture. Traditional CNN style extractors, limited by their local receptive fields, often fail to capture these global characteristics. Consequently, the inpainting frequently lack the distinctive historical aesthetic and material qualities of the original artwork[6].

An optimal inpainting process requires a dynamic allocation of generative focus across different denoising stages, initially prioritizing structure before shifting to style. In contrast, existing methods typically employ static fusion weights. This rigidity can lead to premature style influence disrupting structural formation in early stages. Conversely, it impedes the meticulous refinement of texture in later stages, ultimately compromising the inpainting fidelity to the artistic integrity of the original artwork. Similarly, in other image generation tasks such as low light enhancement, studies have also shown that dynamic guidance is crucial for context enrichment and detail enhancement[7], further highlighting the necessity of introducing dynamic control in mural inpainting.

To overcome the aforementioned limitations, we propose the Decoupled Conditional Adaptive Time dynamic Fusion Diffusion inpainting Method (DCADif). This novel diffusion model framework is specifically tailored for the inpainting of traditional Chinese murals. Its core innovation lies in the fine grained decoupling and dynamic control of a mural structural and stylistic attributes.

The main contributions of this work are as follows:

We propose a Decoupled Condition Encoder that employs parallel pathways to extract distinct representations: structural information from line art and stylistic features from reference images. This architectural separation facilitates independent and precise control over both attributes, thereby providing a robust framework for high-fidelity mural inpainting.

We introduce the SwinStyle Encoder to overcome the inherent limitations of traditional methods in characterizing the complex style of murals. This component is specifically engineered to effectively capture the distinctive historical aesthetic and material qualities of the original artwork.

A Time step-based Adaptive Feature Fusion (TAFF) module is proposed, which prioritizes structural accuracy during the initial stages of denoising and later enhances the inpainting of style and texture, thereby yielding a result that is highly faithful to the original artwork.

The physical inpainting of traditional murals is a highly specialized scientific discipline that demands extensive expertise and technical skill from professional conservators. This process encompasses several key stages, including structural consolidation, surface cleaning, pigment readhesion, filling of lacunae, and inpainting of lost areas. For example, stabilization may involve applying specialized adhesives to consolidate flaking pigment layers. To address lost pictorial areas while preserving historical authenticity, conservators may employ techniques such as 'tratteggio' (inpainting with discernible lines) or 'filling without painting'. However, physical interventions are often irreversible. They also risk introducing the conservator's subjective style, potentially compromising the original artistic intent. Furthermore, in cases of extensive or severe damage, the efficacy of physical inpainting is severely limited.

Advances in digital technology have established non contact digital inpainting as a crucial alternative and supplement to physical methods. Initial digital techniques primarily involved the manual application of tools like the clone stamp by expert operators. While this approach avoided direct physical intervention, it suffered from inefficiency and subjectivity, as the outcome was highly contingent upon the artistic proficiency. These limitations spurred the development of algorithms based on texture synthesis, such as PatchMatch[8], for image inpainting tasks.

Deep generative models have driven significant advancements in image inpainting. Two primary paradigms have emerged: Generative Adversarial Networks and Diffusion Models. GANs compared to earlier methods, are characterized by more sophisticated architectural designs. The LaMa model[9], for instance, utilizes Fast Fourier Convolutions to leverage global contextual information, achieving exceptional performance on large, irregular inpainting tasks. Concurrently, Diffusion Models have become the dominant paradigm, owing to their superior generation quality and training stability. Large-scale, pre-trained models such as the Latent Diffusion Model (LDM)[3] can be adapted for this purpose. When fine-tuned or integrated with modules like ControlNet[10], they can adhere to existing structures while generating highly realistic and diverse content. These advanced technologies offer promising avenues for the inpainting of ancient paintings.

The iterative denoising process of diffusion models enables the generation of highly detailed and photorealistic images. To harness this generative capability for specific tasks, researchers have developed various guidance and control techniques. PnP Diffusion[11], for example, introduced a plug and play method for injecting external feature maps to guide generation. This approach enables flexible control over structure and appearance without necessitating model retraining. Similarly, InstructPix2Pix[12] demonstrated the capacity for complex semantic modifications by enabling image editing via natural language instructions. Such guidance techniques have found direct applications in artistic creation and inpainting. DiffEdit[13], for instance, performs semantic modifications on specific image regions based on textual descriptions, thus offering new approaches for restoring incomplete content.

Despite these significant advances in controllability, the application of such methods to traditional Chinese murals presents considerable challenges. The profound stylistic diversity and compositional complexity of these artworks often lead to a critical trade off. Existing models frequently struggle to reconcile structural fidelity with stylistic consistency, resulting in undesirable artifacts such as style drift or structural distortion.

The pursuit of more precise content control in inpainting has spurred the integration of semantic and stylistic guidance into the generative process. A seminal development in this domain is the Contrastive Language Image Pre training (CLIP) model[14]. CLIP aligns images and text within a shared semantic space through training on vast image-text datasets. This alignment enables CLIP to serve as a powerful semantic guide for image generation and editing. Blended Diffusion[15], for instance, employs CLIP guidance to perform seamless local edits within specified image regions. Specifically for inpainting tasks, CLIP often functions as a perceptual loss that enforces semantic consistency between the restored region and its surrounding context.

Parallel to semantic control, the precise encoding and transfer of artistic style constitutes a central challenge. The decoupling and control of "style" is a broad and active area of research in generative AI. For example, in the task of stylized image captioning, researchers have explored how to use style embeddings to control the specific style of generated text, rather than merely describing image content[16–18]. Such works demonstrate the significant potential of modeling style as a controllable variable, providing valuable insights for a wide range of style-aware generative tasks, including our mural inpainting. Traditional style transfer methods[19] rely on pre-trained VGG networks for style extraction. However, this CNN based approach is often inadequate for capturing the intricate textures and brushwork characteristic of Chinese murals. This limitation motivated a shift towards the Transformer architecture, which excels at modeling long range dependencies. StyTr^2 [20] pioneered the use of a pure Transformer architecture for arbitrary style transfer, and its success is part of a broader trend. The versatility of Transformer based architectures is evident not only in vision tasks but also in complex language generation problems like unified caption summarization[21]. In low level vision inpainting, this success is mirrored by models like Uformer[22], which have demonstrated a superior ability to reconstruct fine-grained textures while preserving global structural integrity. Precisely preserving and reconstructing structural boundaries is a key challenge for high quality image generation, not only in inpainting but also across other computer vision domains. For instance, in medical image segmentation, researchers have proposed ABANet[23], which leverages an attention boundary-aware module to explicitly refine edge features, highlighting a shared pursuit of structural fidelity in cross-domain applications.

Effectively fusing complementary information from different sources is a powerful strategy for enhancing model performance. This idea has been validated in multiple domains; For example, in medical diagnostics, FusionLungNet[24] improves diagnostic accuracy by integrating multi-scale features to effectively capture fine-grained pulmonary details. Current fusion strategies for structure and style information typically employ static weighting. This rigid methodology is misaligned with the intrinsically dynamic nature of the generative process, which progresses hierarchically from coarse structural formation to fine grained textural refinement. The absence of a dynamic, stage aware guidance mechanism therefore represents a critical limitation in current style aware inpainting methods.

Methods

Overall Structure

As illustrated in Fig. 3, our proposed model, DCADif, operates on the principle of decoupled feature extraction and adaptive fusion.Its architecture is designed around three core mechanisms:

Guided Denoising UNet. The fused features are injected into our core UNet to guide the reconstruction.

Decoupled Conditional Encoder. We use a parameter frozen CLIP Sketch Encoder for structure and a bespoke SwinStyle Encoder for style.

Time Step Adaptive Feature Fusion. A dynamic module that adjusts the influence of structure and style based on the denoising timestep.

The following sections will detail each component.

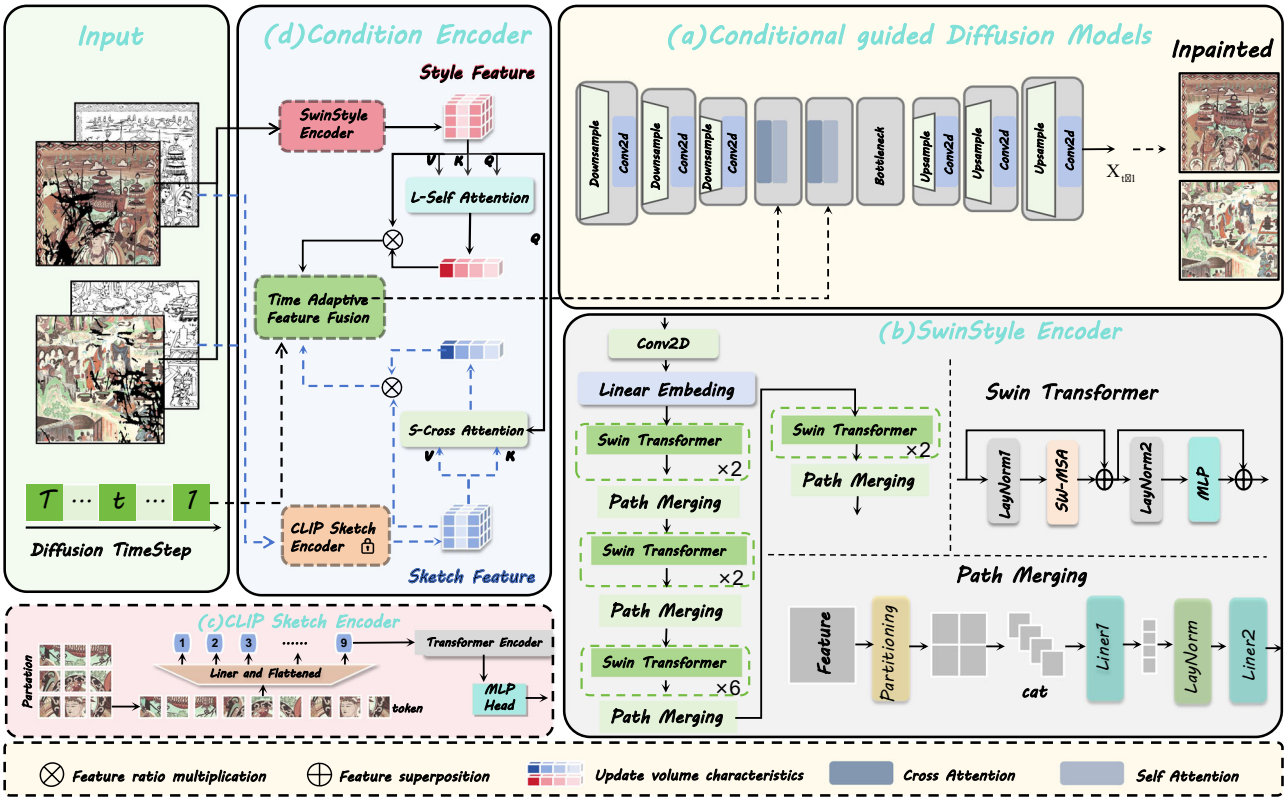

Fig. 3 | Overview of our framework. a Conditional guided Diffusion Models. **b** SwinStyle Encoder. **c** CLIP Sketch Encoder. **d** Condition Encoder.

CLIP Sketch Encoder

To achieve decoupled control over structure and style, we designate line art as the exclusive medium for structural information. Line art inherently disentangles a mural fundamental structure. Its composition, object contours, and spatial layout, from stylistic attributes such as color, lighting, and material texture. We then employ the pre-trained CLIP Sketch Encoder to extract this purified structural representation.

We employ the CLIP image encoder in a parameter frozen, inference only capacity. This process projects the input line art into a high-dimensional latent space, yielding a compact and semantically rich vector we designate as the structural feature $\mathbf{f}_{Sketch}$. Formally, this initial extraction is performed by the pretrained CLIP encoder $\mathcal{E}_{CLIP}$, which processes the line art, I_{line}, to produce an intermediate backbone feature $\mathbf{f}_L$.

$$\mathbf{f}_L = \mathcal{E}_{CLIP}(I_{sketch}) \tag{1}$$

To ensure the structural representation is sensitive to stylistic context, we generate a structural update vector $\mathbf{f}_\Delta$, via a cross attention mechanism. In this operation, the initial structural feature $\mathbf{f}_L$, serves as the Query, while an external style feature $\mathbf{f}_S$, provides the Key and Value. This process, which follows the standard multi head attention mechanism[25], can be summarized as:

$$\mathbf{f}_\Delta = \text{MultiHead}\,(\mathbf{f}_{struct}, \mathbf{f}_{style}, \mathbf{f}_{style}) \tag{2}$$

Finally, the structural update vector $\mathbf{f}_\Delta$ is combined with the initial backbone feature $\mathbf{f}_L$ via element-wise addition. This design, functioning as a residual connection, enables fine-grained, context-aware adjustments while preserving the integrity of the initial structural representation.

$$\left(\mathbf{f}'_{struct}\right)_{i,j,c} = \left(\mathbf{f}_{struct}\right)_{i,j,c} + \left(\mathbf{f}_\Delta\right)_{i,j,c},\ \forall (i,j,c) \in \mathcal{I} \tag{3}$$

This design enables fine grained, context aware adjustments while preserving the integrity of the initial structural representation.

SwinStyle Encoder

To effectively capture the intricate, multi scale stylistic attributes inherent in murals, we employ a Swin Transformer based encoder. This encoder's primary function is to extract a comprehensive and robust style prior from the reference mural image, yielding the style feature vector, $\mathbf{f}_{style}$.

The encoding process begins by converting the input image into a sequence of feature tokens through a standard patch embedding process, which typically involves a convolutional layer followed by a linear projection. These feature tokens are then processed by a multi stage Swin Transformer backbone. This network constructs a hierarchical feature representation through the systematic alternation of Swin Transformer blocks and Patch Merging layers. The Swin Transformer block performs local feature modeling, while the Patch Merging layer downsamples the feature map, thereby expanding the receptive field.

Swin Transformer Block. The ℓ block processes the input features $Z_{\ell-1}$ through the following sequence of operations:

$$\hat{\mathbf{Z}}_l = \text{SMSA}(\text{LN}\,(\mathbf{Z}_{l-1})) + \mathbf{Z}_{l-1} \tag{4}$$

$$\mathbf{Z}_l = \text{MLP}(\text{LN}\,(\hat{\mathbf{Z}}_l)) + \hat{\mathbf{Z}}_l \tag{5}$$

where LN denotes Layer Normalization (LayerNorm), SMSA signifies Shifted Window Multi-head Self-Attention, and MLP is a Multi-Layer Perceptron.

Path Merging. The downsampling operation performed by this layer is formally defined as:

$$\mathbf{f}_{out} = (\mathcal{L}_2 \cdot \text{LN} \cdot \mathcal{L}_1 \cdot \mathcal{R})(\mathbf{f}_{in}) \tag{6}$$

where $\mathbf{f}_{in}$ represents the input feature map, $\mathcal{R}$ denotes the reshaping and concatenation operation, $\mathcal{L}1$ and $\mathcal{L}_2$ are the first and second linear layers, respectively, and LN denotes Layer Normalization.

This hierarchical architecture, characterized by a progressive increase in channel depth and attention heads, enables the encoder to transition its

focus from fine grained textural details in shallow layers to more holistic stylistic patterns in deeper layers. This multi stage process of extraction and abstraction culminates in the generation of a single, highly condensed, and discriminative style vector $\mathbf{f}_S$.

$$\mathbf{f}_S = \text{SwinBackbone}(I_{dam}) \tag{7}$$

where I_{dam} is the damaged mural image serving as the input into the SwinBackbone network.

Finally, to further enhance the representational power, the backbone style feature $\mathbf{f}_S$ is fed into a style self-attention module, L-Self Attention. This module computes a style update vector by identifying and amplifying the most salient patterns within the feature itself. This update vector is then integrated with the backbone feature $\mathbf{f}_S$ via a residual connection, yielding the final, refined style vector, f_{style}. This entire refinement operation is formally defined as:

$$\mathbf{f}_{style} = \mathbf{f}_S + \mathcal{M}_{LSA}(\mathbf{f}_S) \tag{8}$$

where the L-Self Attention module $\mathcal{M}_{LSA}$, is a self-attention mechanism designed to refine the feature representation by allowing its most salient stylistic patterns to interact and reinforce one another. It is defined as:

$$\mathcal{M}_{SA}(X) = Pro\left(Softmax\left(\frac{(XW_Q)(XW_K)^T}{\sqrt{d_k}}\right)(XW_V)\right) \tag{9}$$

where *Pro* denotes the linear projection layer, which maps the features aggregated from the value vectors back to the original C dimensional space to ensure that the output can be residually connected with the input. $\mathbf{f}_s \in R^{N\times C}$ represents the input backbone style feature tensor, where N is the number of *token* and C is the channel dimensionality.

$\mathcal{M}_{LSA}$ denotes the style self-attention module, and $\sqrt{d_k}$ is the scaling factor used to prevent vanishing gradients. $W_Q, W_K \in R^{C\times d_k}$ and $W_V \in R^{C\times d_v}$ are the learnable linear projection weight matrices for generating the query, key, and value, respectively.

Time Step Adaptive Feature Fusion

Conventional conditional diffusion models typically employ static mechanisms, such as cross attention, to integrate external guidance. This rigid approach is fundamentally misaligned with the dynamic nature of the denoising process, where structural guidance is paramount in early stages and stylistic refinement is critical in later ones. Consequently, this misalignment often results in the corruption of structural integrity by premature stylistic influence, leading to significant inaccuracies in the final reconstruction.

To address this challenge, we introduce a novel adaptive fusion mechanism that emulates the coarse to fine strategy of human inpainting experts. This mechanism dynamically modulates the relative influence of structural and stylistic features as a function of the current denoising timestep, t. This ensures that structural guidance dominates in the early, high noise stages, while stylistic and textural refinement prevails in the later, low noise stages.

As illustrated in Fig. 3 (d), the core of the proposed TAFF module lies in generating a set of time dependent weight functions, $t \in (T, \ldots, 1)$ and T. Based on the diffusion model current timestep $\omega_{struct}(t)$ and total number of time steps $\omega_{style}(t)$.

The core of the TAFF module lies in generating a pair of time dependent weights, and, which are computed as a linear function of the normalized timestep :

$$\omega_{struct}(t) = 0.1 + 0.9 \times (t/T) \tag{10}$$

$$\omega_{style}(t) = 0.9 - 0.1 \times (t/T) \tag{11}$$

where T is the total number of diffusion steps. As denoising proceeds, the style weight becomes is completely dominant.

Upon obtaining this set of dynamic weights, we perform a weighted fusion of the final structural feature $\mathbf{f}_{Struct}$ and the style feature $\mathbf{f}_{Style}$ to yield the time dependent fused conditional feature.

$$\mathbf{f}_{fu}(t) = \left[\mathbf{f}_{struct}, \quad \mathbf{f}_{style}\right] \cdot \mathbf{w}(t) \tag{12}$$

where w is the time dependent weight vector $\mathbf{w}(t) = [\omega_{struct}(t), \omega_{style}(t)]^T$.

This fused feature $\mathbf{f}_{fu}(t)$ is subsequently injected into the UNet bottleneck layer of the diffusion model to provide dynamic guidance for each denoising step.

Within the UNet denoising process at each time step t, the input noisy latent X_t, is processed by the downsampling path to produce a bottleneck feature representation $\mathbf{h}_{bo}$. This bottleneck feature is then modulated by our time adaptive fused feature $\mathbf{f}_{fu}(t)$. To ensure dimensional compatibility, $\mathbf{f}_{fu}(t)$ is first passed through a linear projection layer ϕ to match its dimensionality and then inject it into the network via feature addition to guide the generation process. This guidance step can be formulated as:

$$\mathbf{h}'_{bo}(\mathbf{X}_t, t) = \mathbf{h}_{bo}(\mathbf{X}_t, t) + \phi(\mathbf{f}_{fu}(t)) \tag{13}$$

where $\mathbf{h}_{bo}(\mathbf{X}_t, t)$ is the original feature extracted from the noisy image X_t by the U-Net bottleneck layer at timestep t. ϕ is a lightweight projection network for feature alignment, and $\mathbf{h}'_{\text{bo}}(\mathbf{X}_t, t)$ is the updated bottleneck feature after dynamic conditional guidance.

This strategic temporal decoupling facilitates the reconciliation of macro-structural integrity with fine-grained stylistic details. The result is a reconstruction that achieves a superior degree of fidelity to the original artwork.

Loss Function

The training is governed by a composite objective function that integrates several loss components to jointly optimize for both structural fidelity and stylistic realism.

L1 Loss. It is particularly effective at preserving high frequency structural details, such as edges and contours. This effectiveness stems from its superior robustness to outliers compared to other pixel level losses. The formal is as follows:

$$L_1 = \frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i)| \tag{14}$$

where y_i denotes the ground truth image, and $f(x_i$ is the reconstructed output image.

MSE Loss. It evaluates error by computing the sum of the squared differences between predicted and ground truth values. However, the quadratic nature of this penalty renders MSE highly sensitive to large pixel deviations, which in practice often leads to overly smoothed results that lack the fine textures crucial to artworks. It is formulated as:

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \tag{15}$$

Preceptual Loss. It does not perform a direct comparison in pixel space. It computes the feature distance between image patches within the feature space of a pre-trained VGGNet. It can be summarized as:

$$L_{Preceptual} = \frac{1}{n}\sum_{i=1}^{n}(F_i(x) - F_i(y))^2 \tag{16}$$

where x is the input image and y is the target image, $F_i(x)$ and $F_i(y)$ respectively denote their feature representations at the i layer of a pretrained neural network, and N is the number of feature layers.

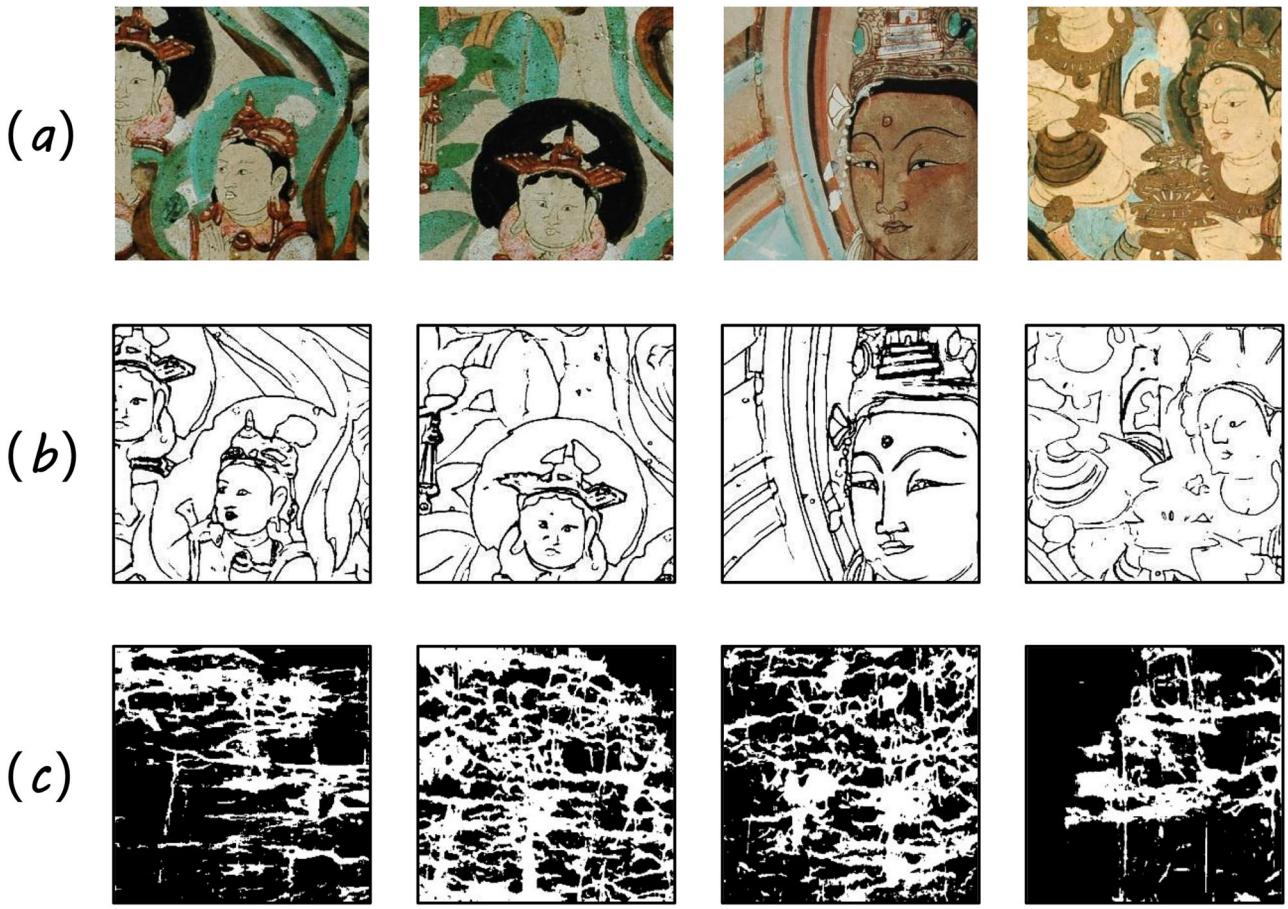

Fig. 4 | Examples of Mural paintings. a is Dunhuang murals. **b** is the line sketch of the Dunhuang murals. **c** is the real mask. Unlike commonly used synthetic masks, these masks realistically replicate the complex patterns of cracks, fading, and pigment loss that occur over time. Training with such real-world degradation patterns enables our model to generalize more effectively to authentic mural restoration scenarios.

LPIPS Loss. The Learned Perceptual Image Patch Similarity loss is designed to more accurately reflect human perceptual judgment than traditional perceptual losses. It calculates the distance between deep features of two images, weighted by learned linear layers to better match human perception. The loss is computed as:

$$L_{LPIPS} = \sum_{l} W_l \left| F_l(x) - F_l(y) \right|_2^2 \tag{17}$$

where x is the generated image and y is the target image,$F_l(x)$ and $F_l(y)$ are the unit normalized feature representations extracted from the $l-th$ layer of a pretrained network, and W_l is a learned weight vector used to scale the contribution of each layer's feature distance.

Total Loss. It is represented as:

$$\begin{aligned} L_{total} &= \lambda_1 L_{1_n} + \lambda_2 L_{MSE_n} \\ &+ \lambda_3 L_{1_i} + \lambda_4 L_{Pre_i} + \lambda_5 L_{LPIPS} \end{aligned} \tag{18}$$

where L_{1_n}, L_{MSE_n} represent the L1 and Mean Squared Error losses calculated between the predicted noise and the ground truth noise, respectively. While L_{1_i}, L_{Pre_i} enote the L1 and perceptual losses computed between the denoised image and the ground truth image. The weight parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 are weight parameters set to 0.5, 0.5, 0.5, 1, 0.1, respectively.

Datasets

MuralVerse-S. We propose a dataset of murals from various regions of China, comprising 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.

The dataset was curated from images procured from collaborating institutions and digital art databases. The curation process involved a rigorous screening and classification performed by professional artists. Artworks were categorized based on their distinct styles, dynasties, and color palettes to ensure the final dataset diversity and representativeness.

The data preparation pipeline for each mural initiates with the extraction of its corresponding line art. Subsequently, natural damage is simulated by manually applying masks to the intact ground truth image. This process yields a complete training sample, consisting of the damaged image, the binary mask, and the structural line art, which collectively serve as the input for network training.

MaskCLP-S. The dataset is obtained from relevant cooperative research institutions. It comprises 8273 images of Chinese landscape paintings, as illustrated in Fig. 5. The dataset is divided into 7446 training images and 827 testing images, all cropped to 256 × 256. This dataset encompasses a wide range of traditional paintings from various historical dynasties, featuring the unique styles of numerous outstanding artists.

Ethics Statement

The dataset used in this study is publicly available and has received the necessary approval for use. All images, videos, and associated personal information are published in accordance with the licensing terms of the dataset, and the researchers have adhered to the terms provided by the dataset's publisher. Since the dataset is publicly accessible and includes

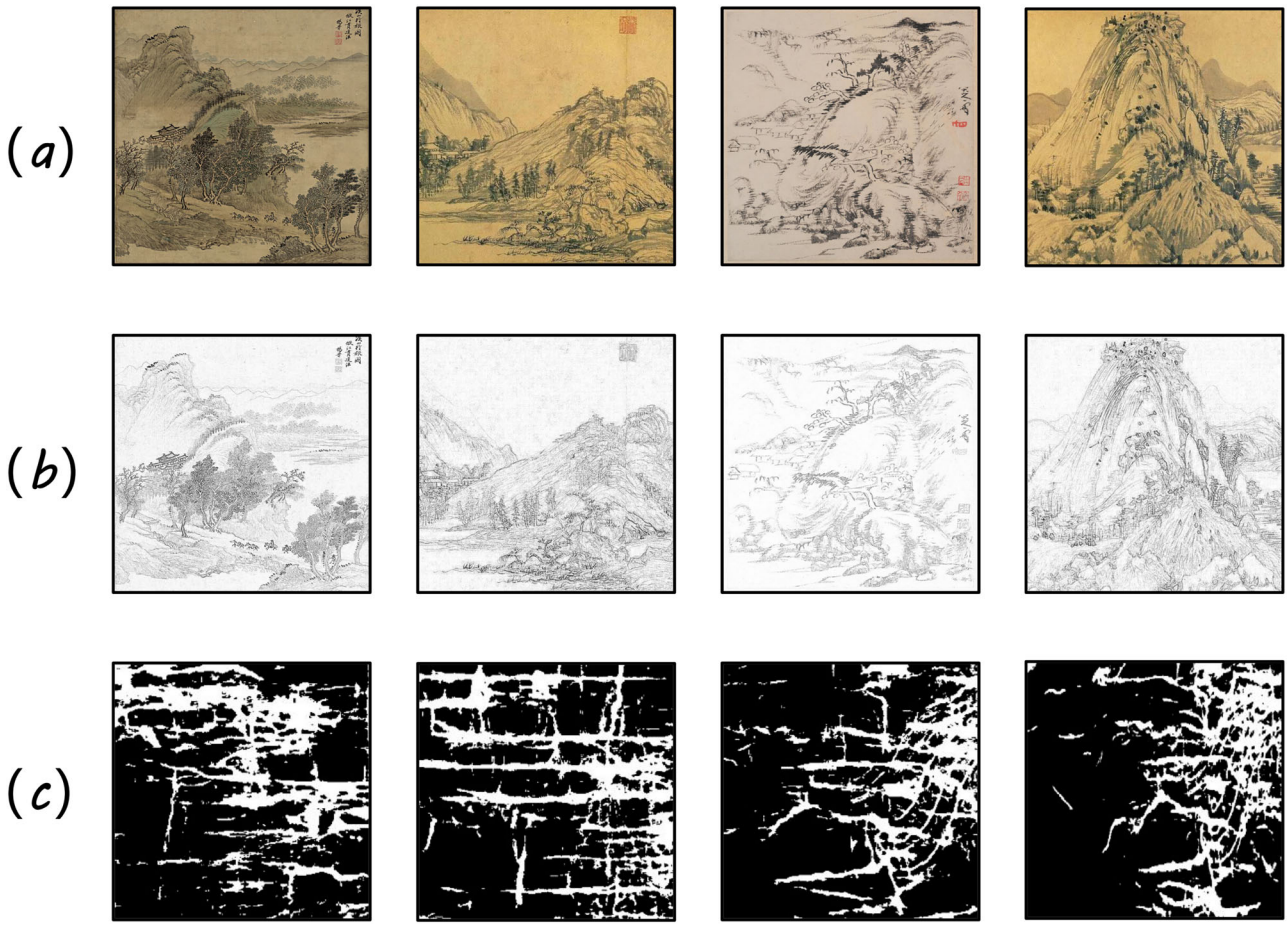

Fig. 5 | Demonstrating the generalization capability of our model on the related domain of Chinese landscape paintings. Row **a** presents examples from various *Qing* Dynasty landscape paintings, including works by artists such as *Wang Jian*, *Hua Yan* and so on. Although these paintings differ from murals in medium and brushwork techniques, they share a common emphasis on linear structure and stylistic mood. Row **b** contains the corresponding line art extracted for structural guidance, and Row **c** shows the degradation masks applied for testing.

content with the required authorization, we confirm that the individuals involved have provided consent at the time of dataset publication.

Implementation Details

In our experiments, models implemented with PyTorch were trained on NVIDIA H20 GPUs. Prior to training, we employed a series of data augmentation techniques to enhance model performance and robustness. These techniques include resizing, cropping, rotation, flipping, and noise addition. Original painting images were resized to a uniform resolution of 256 × 256. The batch size was set to 32. The models were trained for 2000 epochs. A dynamic learning rate schedule was utilized, which progressively annealed the learning rate throughout the training process to ensure stable convergence.

To rigorously evaluate the model's generalization ability to unseen artistic styles, we carefully partitioned our dataset. Specifically, we ensured that artworks from the same dynasty or by the same artist did not simultaneously appear in both the training and testing sets. This partitioning strategy, analogous to a leave one dynasty out cross validation, is designed to prevent the model from achieving high scores simply by memorizing specific styles.

Model Complexity and Efficiency

The proposed DCADif is a large scale diffusion model, reflected in its complexity metrics. It comprises a total of 561.68 million parameters, positioning it as a substantial network designed to capture intricate artistic features. The computational cost for a single denoising step on a 256 × 256 input is 486.14 GFLOPs. As a diffusion model, the final image generation is an iterative process. The total end to end inference time, which includes the full sampling loop, was benchmarked on a single NVIDIA H20 GPU at 377.00 ms per image. This corresponds to a throughput of approximately 2.65 FPS. While computationally intensive, this scale is crucial for achieving the high fidelity restoration results demonstrated in our experiments.

Evaluation Metrics

We quantitatively assess the inpainting quality using three standard metrics: PSNR[26], SSIM[27], and LPIPS[28]. PSNR and SSIM quantify pixel level fidelity and structural correspondence, respectively, while LPIPS evaluates perceptual realism by measuring similarity in a deep feature space. We quantitatively assess the inpainting quality using four standard metrics: PSNR[26], SSIM[27], LPIPS[28], and Gram-matrix style loss[19]. PSNR and SSIM quantify pixel-level fidelity and structural correspondence, respectively, while LPIPS evaluates perceptual realism and Gram-matrix style loss specifically assesses artistic style fidelity.

Results

Baselines

For a comprehensive performance evaluation, we benchmark our model against nine representative baseline methods, which are grouped into three distinct categories:

(1) CNN. These methods employ CNN architecture for image inpainting.

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

CTSDG[30]: A coupled texture structure decomposition network implementing dual-stream inpainting through task-specific subnetworks.

RFR[31]: A progressive inpainting framework employing iterative refinement with cascaded recurrent feedback modules.

EC[32]: A CNN based line art colorization model that uses adaptive normalization to inject structural prior, yielding high-fidelity colorization without diffusion steps.

(2) Transformer. These methods employ Transformer architecture for image inpainting.

MAT[33]: The pioneering transformer based large hole inpainting framework combining global attention mechanisms with local convolutional features for high resolution inpainting.

PromptIR[34]: A prompt driven transformer that unifies all image inpainting tasks via textual degradation queries, dispensing with task specific branches.

(3) Diffusion. These methods employ diffusion model architecture for image inpainting.

Strdiffusion[35]: A lightweight diffusion sampler with momentum based skip for fast, multi scale inpainting.

SDE[36]: Score based generative framework using stochastic differential equations for high-quality image synthesis and inpainting.

RePaint[37]: A diffusion inpainting method that couples denoising sampling with mask-aware reverse SDE for structure and texture consistency.

Quantitative Analysis

To validate the efficacy of the proposed DCADif framework, we conducted a comprehensive benchmark against leading state-of-the-art methods using our proprietary MuralVerse-S dataset. This evaluation encompassed both quantitative metrics and qualitative visual comparisons.

We conducted a comprehensive quantitative benchmark of the proposed DCADif model against a diverse set of leading image inpainting methods. This selection intentionally spans multiple architectural paradigms, including CNN based, Transformer based, and Diffusion based approaches.

Table 1 summarizes the quantitative results. To guarantee a rigorous and unbiased comparison, all baseline methods were retrained from scratch on our proprietary MuralVerse-S dataset. The models were evaluated across three distinct ranges of random mask ratios (0.1-10%, 10-20%, and 20-30%) to assess their performance under varying levels of degradation.

In terms of PSNR, DCADif consistently outperforms all baseline methods. This performance advantage becomes increasingly pronounced at higher mask ratios, demonstrating the superior robustness to severe degradation. This trend is best illustrated by the comparison with its leading Diffusion based competitor, RePaint. In the most challenging scenario, 20–30% mask ratio, DCADif achieves a performance margin of 0.51 dB.

DCADif demonstrates exceptional performance stability on the SSIM metric, particularly as the degree of image degradation increases. This stability stands in sharp contrast to methods like EC, which exhibit a performance degradation of over 25% at high mask ratios, indicating a critical failure in structural reconstruction. Furthermore, DCADif surpasses the similarly robust RePaint model, achieving a 3.6% relative performance gain under high mask ratios. This margin underscores its superior capacity for maintaining global structural consistency.

DCADif demonstrates a commanding lead on the LPIPS metric, indicating a substantial improvement in perceptual realism. Relative to the second-best performing model, RePaint, DCADif reduces perceptual error by a remarkable 50% to 57%. Notably, this performance gap widens as the level of image degradation increases.

Moreover, its exceptional performance stability across varying levels of inpainting difficulty validates the sophistication and efficiency of our model's design, establishing it as a robust and reliable new benchmark in the field of image inpainting Fig. 6.

Table 1 | Comparison results on DCADif

Model	Type	PSNR↑			SSIM↑			LPIPS↓			Gram↓		
		0.1–10%	10–20%	20–30%	0.1–10%	10–20%	20–30%	0.1–10%	10–20%	20–30%	0.1–10%	10–20%	20–30%
CTSDG[30]	GAN	26.449	26.113	26.016	0.782	0.774	0.771	0.271	0.275	0.278	0.000046	0.000096	0.000199
AdaIR[29]	CNN	23.585	22.562	21.268	0.917	0.875	0.853	0.082	0.119	0.141	0.001535	0.001535	0.001871
EC[32]	GAN	25.907	20.183	18.455	0.933	0.741	0.699	0.047	0.201	0.231	0.000098	0.000842	0.001171
RFR[31]	CNN	23.496	18.571	18.617	0.891	0.654	0.648	0.122	0.311	0.302	0.000363	0.002007	0.001639
PromptIR[34]	Transformer	21.164	20.675	20.393	0.555	0.554	0.551	0.393	0.391	0.394	0.000796	0.00108	0.001377
MAT[33]	Transformer+GAN	25.892	25.591	25.479	0.743	0.737	0.733	0.293	0.297	0.301	0.000058	0.000165	0.00031
SDE[36]	Diffusion	22.583	18.819	16.182	0.930	0.772	0.706	0.059	0.170	0.226	0.000071	0.00038	0.00063
Strdiffusion[35]	Diffusion	25.826	25.622	25.021	0.898	0.881	0.873	0.042	0.042	0.045	0.000043	0.000055	0.000064
RePaint[37]	Diffusion	27.474	26.981	26.433	**0.931**	0.901	0.893	0.026	0.029	0.030	0.000043	0.000053	0.000061
Ours	Diffusion	**27.798**	**27.262**	**26.942**	0.926	**0.926**	**0.925**	**0.013**	**0.013**	**0.013**	**0.00004**	**0.000049**	**0.000053**

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.

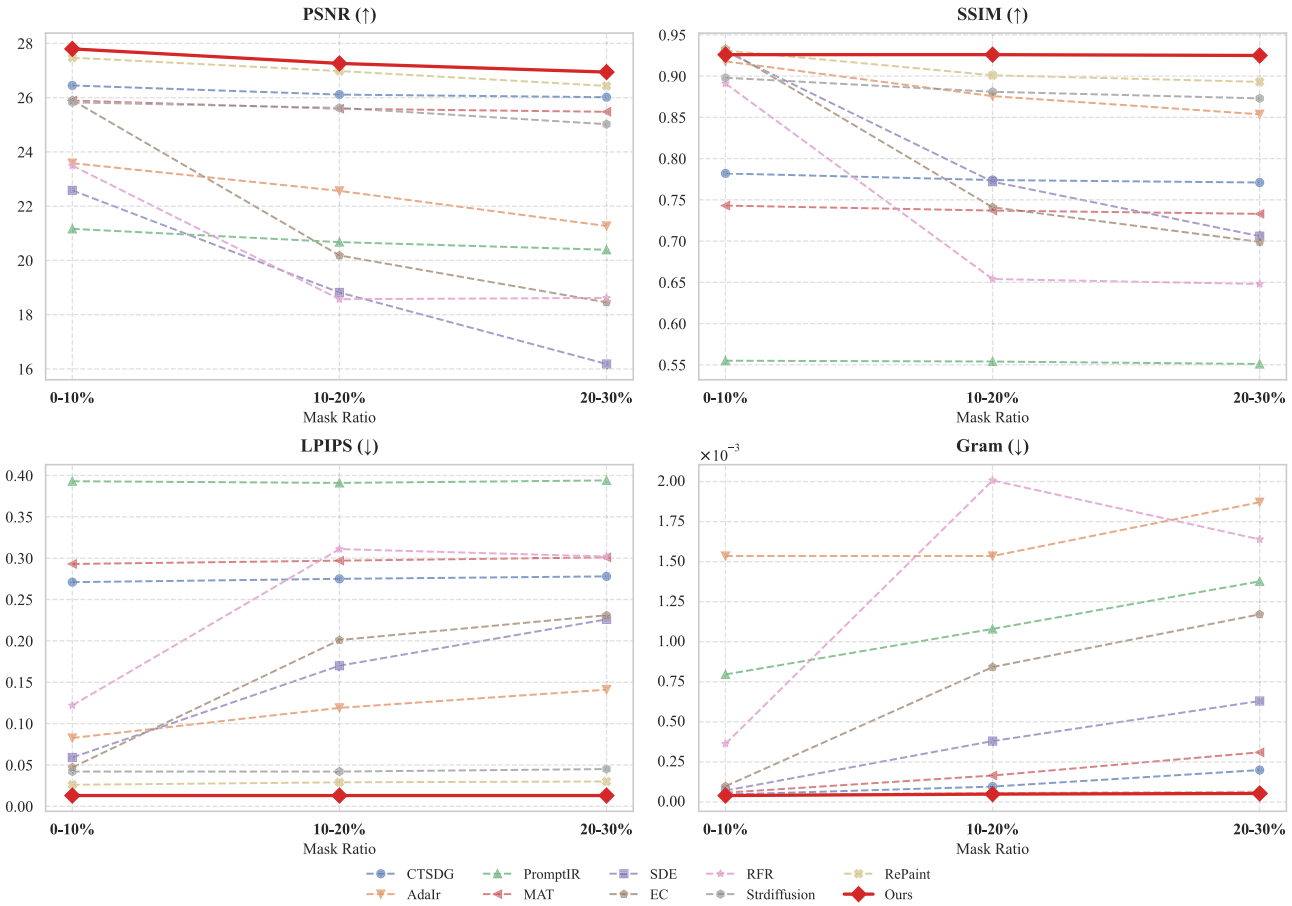

Fig. 6 | A visualization of the quantitative comparison results from Table 1. These line plots track the performance trends of our proposed model (Ours) against baseline methods across varying mask ratios. Each subplot corresponds to one of four metrics: PSNR, SSIM, LPIPS, and Gram matrix style loss.

Qualitative Analysis

As shown in the comparative results in Figs. 7 and 8, our model demonstrates a superior capability to generate visually plausible and high-fidelity results, markedly outperforming other mainstream methods.

The task in Fig. 8a requires the reconstruction of significant facial and sartorial details. The magnified inset (a1) highlights the limitations of the baseline methods. CTSDG and MAT, for instance, generate overly smoothed results that fail to recover crucial facial features or the hat's texture. RFR exhibits catastrophic failure, producing incoherent artifacts with no semantic relevance to a face. While the result from RePaint is plausible, it lacks the requisite sharpness and fine detail of the ground truth (GT). In contrast, DCADif successfully reconstructs the face with well defined features and contours. It also restores the hat's intricate texture, yielding a final inpainting nearly indistinguishable from the ground truth.

In Fig. 8b showcases the inpainting of an ornate headdress, a task defined by its intricate details and fine linework. The magnified inset (b1) underscores the difficulty of preserving such high frequency details. RFR again fails to generate coherent content, producing chaotic artifacts and demonstrating an inability to model the image underlying structure. CTSDG and MAT render the fine lines as an indistinct blur, thereby compromising the design structural integrity. Although RePaint captures the overall form, it fails to reproduce the linework with sufficient sharpness, resulting in a loss of definition. In contrast, our model performs exceptionally well, accurately reconstructing both the fine linework and the subtle background color gradations.

These visual comparisons provide empirical validation for our quantitative results. While competing methods often suffer from blurring, artifacting, and structural inconsistencies, DCADif consistently delivers inpaintings that are semantically coherent, rich in detail, and stylistically faithful to the original artwork.

Comparision with Celebrated Tanditional Mural Painting

To further assess its generalization and practical utility, the DCADif framework was applied to the inpainting of authentic ancient murals exhibiting natural degradation. Unlike the synthetic masks used for training, these cases feature complex, compound forms of degradation, including color fading, pigment flaking, and structural cracks, posing a formidable test of the robustness. The successful outcomes of these inpainting underscore the substantial potential of DCADif as a viable tool for digital cultural heritage preservation.

In Fig. 9, we present the inpainting results for three cases of authentic murals. The versatility and robustness are evident in its handling of diverse degradation types. It successfully addresses the diffuse mottled stains of 'the Winged Beast mural', reconnects the fractured structural lines of 'the Charging Bull mural', and reconstructs the holistically deteriorated scene of 'the Court Ladies and Apsaras mural'. This performance on authentic artifacts validates the efficacy of DCADif as a practical inpainting tool. Furthermore, this success underscores the significant potential for contributions to digital archaeology, museology, and the broader field of cultural heritage preservation.

Comparision on Diverse Datasets

To further validate our model, we conducted additional experiments on the MaskCLP-S dataset. This dataset encompasses a wide range of traditional paintings from various historical dynasties, featuring the unique styles of numerous outstanding artists.

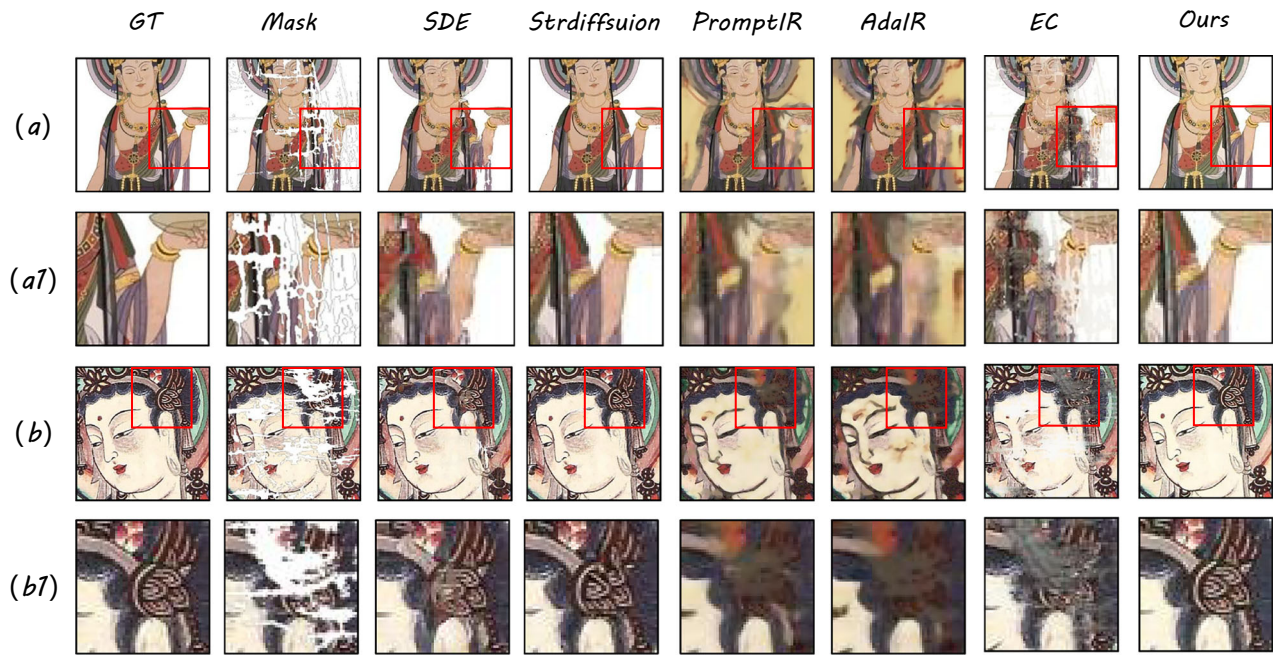

Fig. 7 | Qualitative comparison with state of the art methods, including SDE, Strdiffusion, PromptIR, AdaIR, and EC. The figure presents restoration results for two different examples: (**a**) The first example, and (**b**) the second example. Rows (a1) and (b1) provide zoomed in views of the regions highlighted by red boxes for detailed comparison. Note that baseline methods often suffer from artifacts, blurriness, or stylistic inconsistencies. In contrast, our DCADif successfully reconstructs both fine textures and clean contours with high fidelity to the original artwork in both examples.

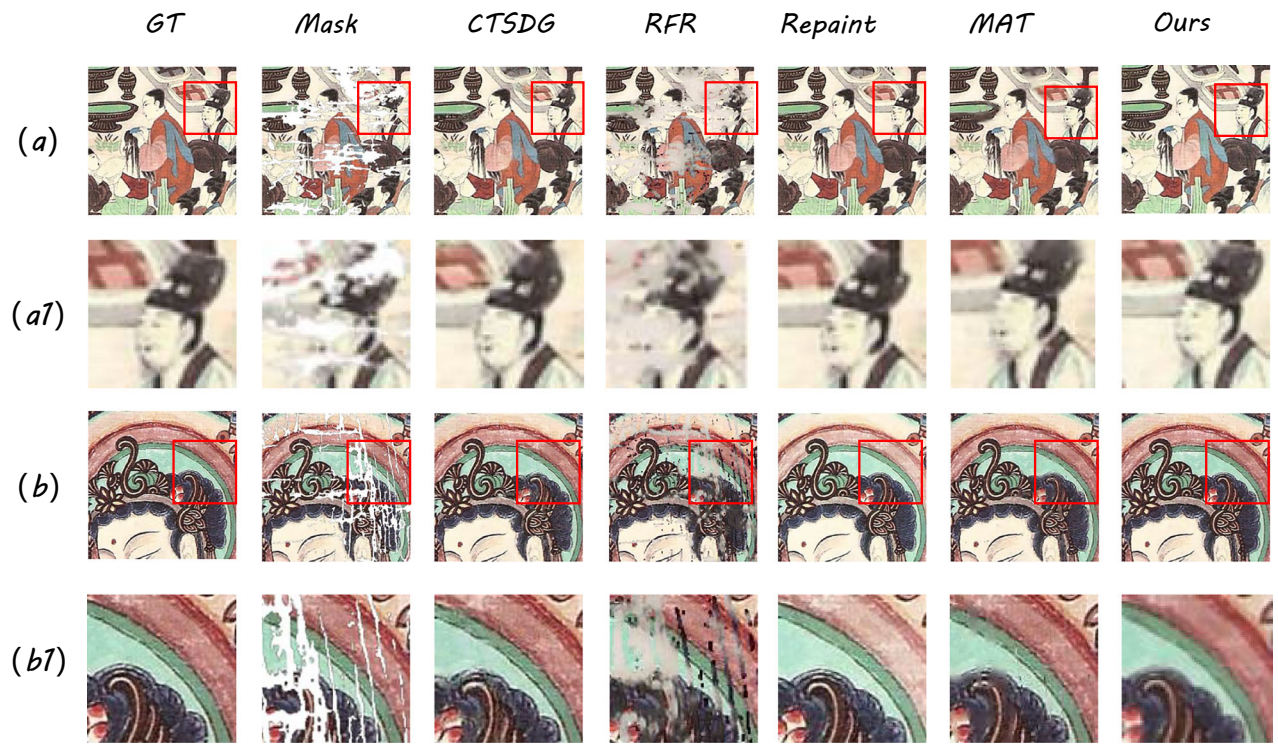

Fig. 8 | Qualitative comparison with state of the art methods, including SDE, Strdiffusion, PromptIR, AdaIR, and EC. The figure presents restoration results for two different examples: (**a**) The first example, and (**b**) the second example. Rows (a1) and (b1) provide zoomed in views of the regions highlighted by red boxes for detailed comparison. Note that baseline methods often suffer from artifacts, blurriness, or stylistic inconsistencies. In contrast, our DCADif successfully reconstructs both fine textures and clean contours with high fidelity to the original artwork in both examples.

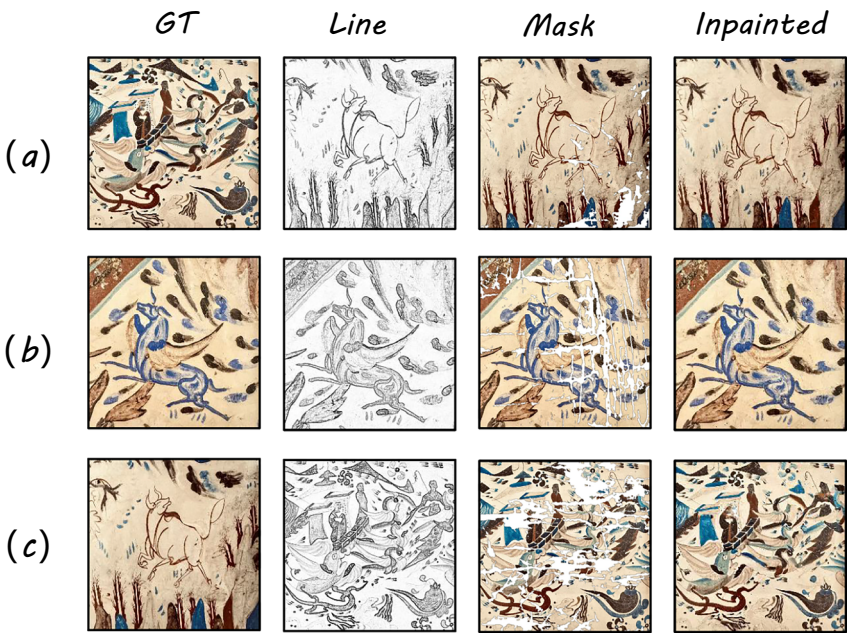

Fig. 9 | Comparison on Traditional Chinese Painting. Comparison on Traditional Chinese Painting. The figure displays results from three different examples, shown in rows (**a**), (**b**), and (**c**). 'GT' denotes the ground truth image. 'Line' represents the corresponding line art. 'Mask' indicates the synthetically damaged painting, and 'Inpainted' shows the restored image.

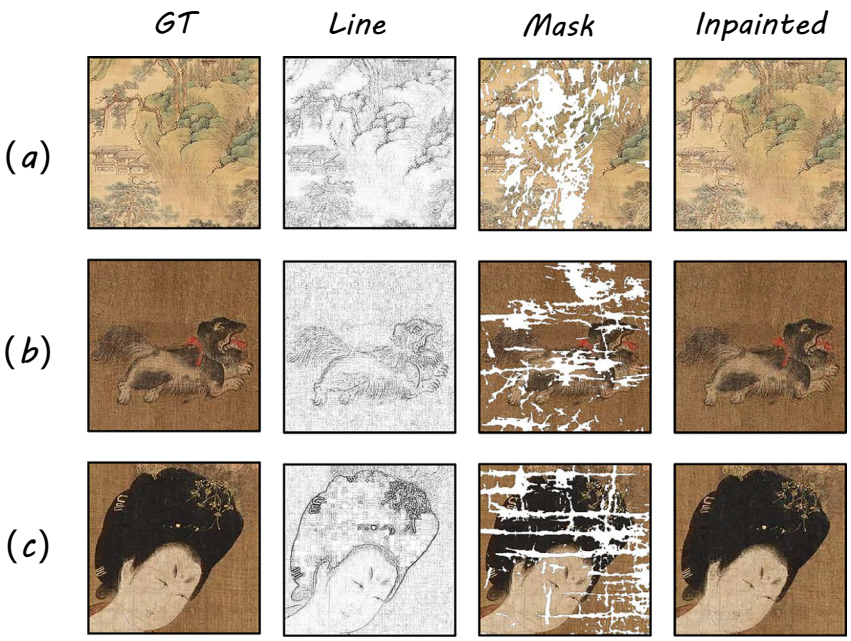

Fig. 10 | Comparison on Traditional Chinese Painting. Comparison on Traditional Chinese Painting. The figure displays results from three different examples, shown in rows (**a**), (**b**), and (**c**). 'GT' denotes the ground truth image. 'Line' represents the corresponding line art. 'Mask' indicates the synthetically damaged painting, and 'Inpainted' shows the restored image.

The experimental results demonstrate that, during the pixel wise decoding of missing regions, the model not only aligns the local brushwork with the style of the original artwork but also, through its adaptive fusion strategy, ensures that the ink tones create a natural transition with the surrounding strokes. As shown in Fig. 10, in a landscape painting by *Wang Shigu* where approximately 18% of the left mountain ridge was damaged by mildew, the inpainting result not only recovers the fine layers of the 'hemp fiber strokes' but also precisely reproduces the texture of the dry brush work, rendering the repaired boundary virtually imperceptible.

Comparision with Celebrated Traditional Chinese Painting

Digital inpainting must strike a balance between semantics and aesthetics. It must not only precisely restore the physical structure and stylistic characteristics of missing areas but, more importantly, ensure that the result integrates seamlessly with the original artwork. By incorporating a multi level style perception mechanism and global context modeling, our model effectively captures the subtle continuity of artistic intent across broken brushstrokes a key characteristic of traditional painting thereby achieving a new equilibrium between visual realism and aesthetics. For our experiments, we cropped sections from the Tang dynasty painting *Court Ladies Wearing Flowered Headdresses* and a landscape painting by *Wang Shigu* to create the images for inpainting. Figure 10 presents the qualitative inpainting results of our proposed model on the renowned painting *Court Ladies Wearing Flowered Headdresses*, with source data details provided in Fig. 11. The experimental results demonstrate that our model achieves a remarkable balance between semantic coherence and aesthetic fidelity. The model transcends basic pixel filling to perform context aware semantic inpainting. For instance, in subplot (b), the model not only recovers the correct color of the dog's coat but also precisely reconstructs its

Fig. 11 | The image is One of Top Ten China Famous Paintings. This is the artwork *Court Ladies Wearing Flowered Headdresses*, painted by the *Tang* Dynasty artist *Zhou Fang*. The red boxes indicate two representative regions chosen for our inpainting experiments: the lady's face with her high chignon, and the small dog below. The two panels on the right are the corresponding magnified views of these regions, which serve as the Ground Truth for our experiments. The unique style and fine details of this artwork present a challenging scenario for validating the capabilities of our model.

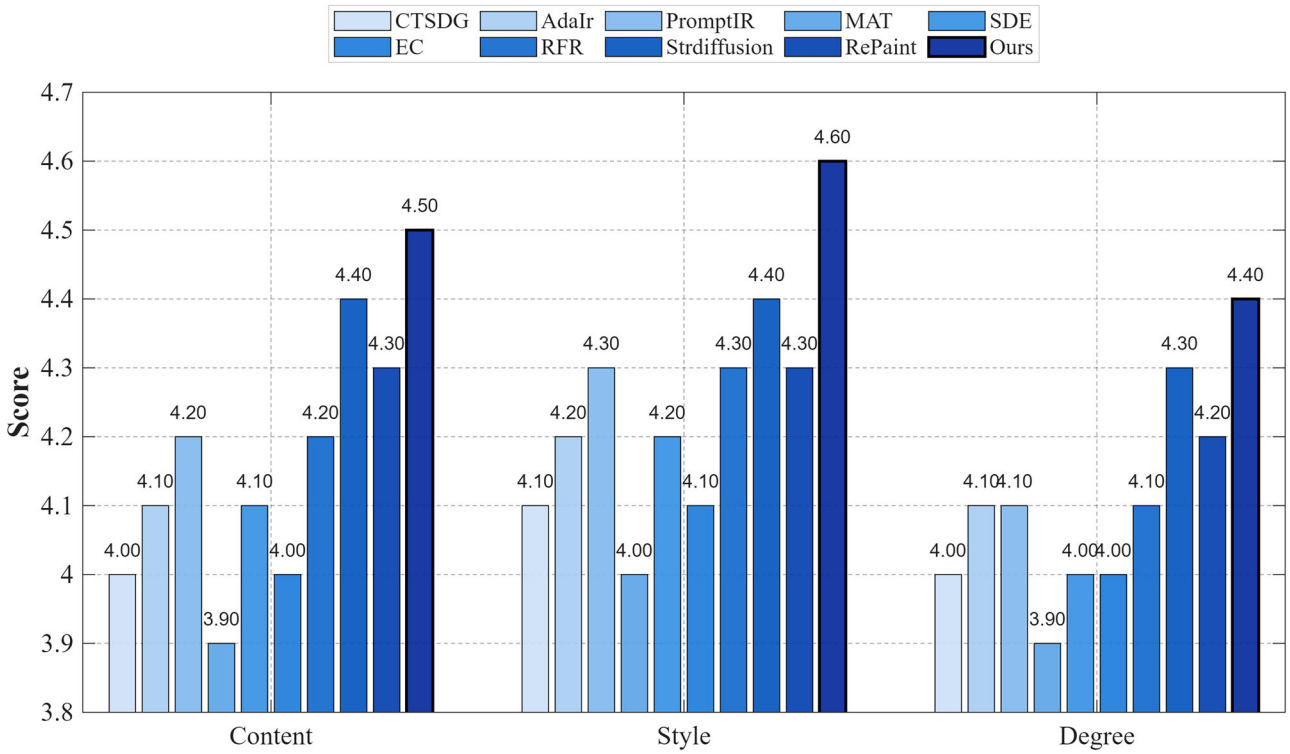

Fig. 12 | Illustration of the User Study.

directional flow, maintaining fine grained textural consistency. Furthermore, subplot (c) highlights the exceptional ability to preserve structural integrity, as evidenced by the high fidelity reconstruction of facial contours and edges. Collectively, these results demonstrate that our model is capable of understanding and generating content that is both visually plausible and semantically meaningful within the context of classical Chinese painting.

User Study

We recruited 50 participants, including faculty and graduate students specializing in art. We used several competing models, as well as our own, to restore a set of murals and presented the resulting images to the participants. They were then asked to rate the results based on the following three criteria: (a) Content Consistency: the degree to which the content of the restored image is consistent with the original. (b) Style Fidelity: the extent to which the brushwork, color, and texture reproduce those of the original mural. and (c) Degree: a comprehensive assessment of how well the mural was restored. As shown in Fig. 12, DCADif demonstrates outstanding performance.

A rating scale from 0 (Worst) to 5 (Best) was used for each criterion, where a higher score indicates a more favorable evaluation. The rating scale was defined as follows:

$$\text{Score} = \frac{\sum_{i=1}^{n}(f_i \cdot w_i)}{P} \tag{19}$$

where P is the number of participants who answered the question, f_i denotes the frequency of the i-th option being selected, and w_i represents the weight of the i-th option determined by its ranking.

Effectiveness of Componets

To systematically deconstruct the DCADif model and validate its architectural design, we conducted a key ablation study focused on the selection of the Style Encoder and the Sketch Encoder. We evaluated four different combinations of two powerful encoders for extracting style and structure information: CLIP, for its high-level semantic understanding, and SwinStyle, for its ability to capture both local and global visual features. The qualitative results are shown in Fig. 13. The quantitative results presented in Table 2.

The experimental results in Fig. 13 demonstrate that the heterogeneous combination, employing SwinStyle as the style encoder and CLIP as the sketch encoder achieved decisively superior performance. This

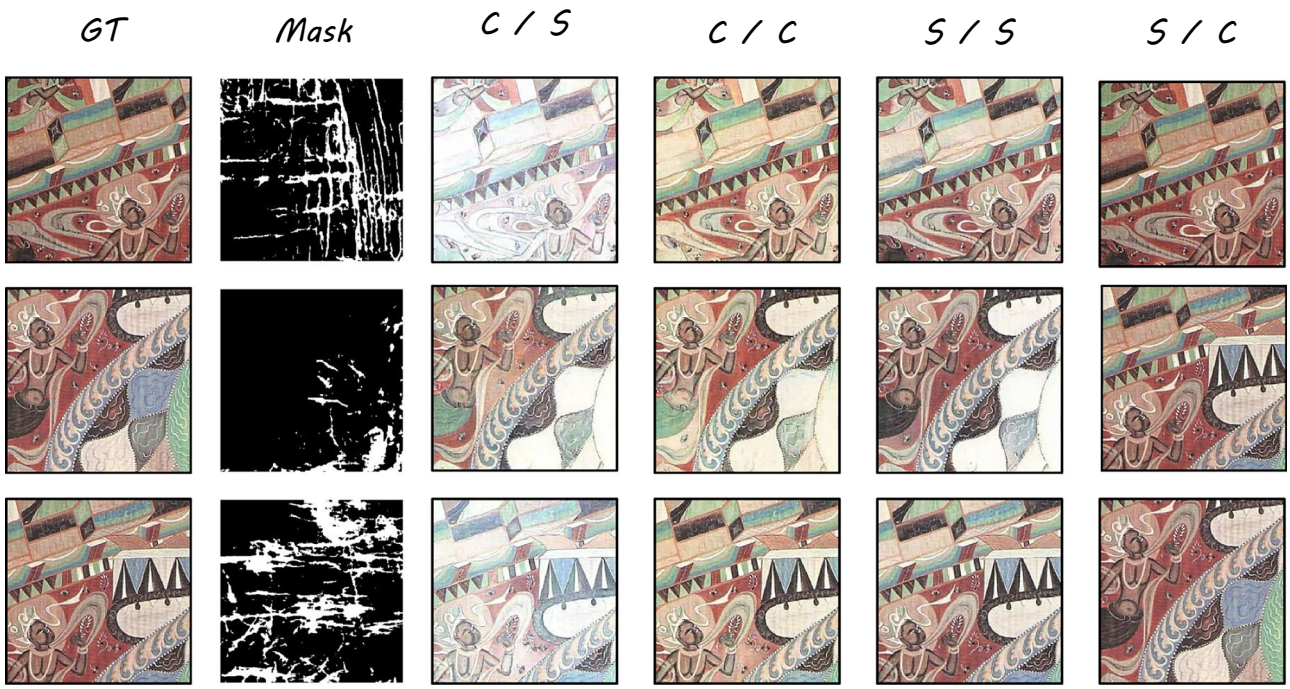

Fig. 13 | Ablation Study of Encoder-Decoder. Visual ablation study of our encoder design. `GT' denotes the ground truth and `Mask' is the damaged input. The other columns compare different encoder combinations for style and line art extraction, formatted as [Style]/[Line Art], using either a CLIP encoder (C) or our SwinStyle encoder (S).

Table 2 | Ablation study on the Style Encoder and Sketch Encoder within the DCADif framework

Style/Sketch	PSNR↑	SSIM↑	LPIPS↓
SwinStyle/CLIP	21.887	0.821	0.043
CLIP/CLIP	21.455	0.875	0.028
SwinStyle/SwinStyle	24.596	0.922	0.019
CLIP/SwinStyle	**26.942**	**0.925**	**0.013**

In the first column of the table, the encoder configurations are specified as (Style Encoder / Sketch Encoder). The ablation study was conducted with a mask ratio of 20%-30%. 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**.

Table 3 | Ablation study on the structural encoder

CLIP Model	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Frozen CLIP	26.942	**0.926**	0.013
Frozen CLIP (Ours)	**26.942**	0.925	**0.013**

We compare the performance with a frozen CLIP encoder versus without it (W/O frozen CLIP). The test was conducted on masks with a 20%-30% corruption ratio. *Optimal results are displayed in **bold**.

configuration not only attains the best results across all three key metrics but, more significantly, a comparison with the other combinations provides profound insight into the intrinsic logic.

First, when we used SwinStyle to extract style and CLIP to extract the sketch, a catastrophic decline in performance was observed: the PSNR plummeted by over 5.0 dB, and both SSIM and LPIPS deteriorated substantially. This provides compelling evidence that SwinStyle expertise in precisely parsing the local structure and edge information of line art is irreplaceable by CLIP, while concurrently demonstrating that CLIP capacity for capturing and encoding high level, abstract style semantics far surpasses that of SwinStyle. This finding clearly delineates the 'capability boundaries' of the two encoders, confirming that their assigned roles are both correct and uniquely suited.

Second, the performance of the homogeneous combinations further reinforces our design rationale. When two CLIP encoders were used, the model demonstrated the poorest performance in structural reconstruction despite its stylistic understanding, registering one of the lowest SSIM scores among all combinations. This underscores CLIP shortcomings in fine grained structural perception. Conversely, when two SwinStyle encoders were used, the model excelled on the SSIM metric, performing nearly on par with the optimal configuration, but showed a noticeable gap in PSNR and LPIPS. This indicates that while SwinStyle is highly effective at processing structural information, it lacks CLIP ability to associate image content with high level semantic style, resulting in generated textures and details that are less rich and realistic.

In conclusion, CLIP powerful semantic understanding makes it the ideal choice for extracting style information, while SwinStyle fine grained perception of visual patterns makes it most effective for parsing structural contours. The ability of DCADif to synergistically process these two distinct information streams is what enables it to achieve state of the art performance in image inpainting tasks.

Effectiveness of Frozen CLIP for Structural Guidance

To justify the use of a frozen CLIP encoder for structure extraction, we balanced the trade-off between fine-tuning for domain adaptation and freezing the weights to preserve robust, general-purpose features. While fine-tuning may enhance specialization, it carries the risk of catastrophic forgetting of the extensive knowledge acquired during large-scale pre-training. To validate this design choice, we conducted an ablation study comparing our full model against a baseline variant devoid of the CLIP encoder. As demonstrated in Table 3, the quantitative results highlight the significant contribution of the frozen CLIP guidance.

The results of the ablation study demonstrate the nuanced role of the frozen CLIP encoder. Regarding pixel-level fidelity, both configurations achieved identical performance, and the perceptual quality (LPIPS) showed

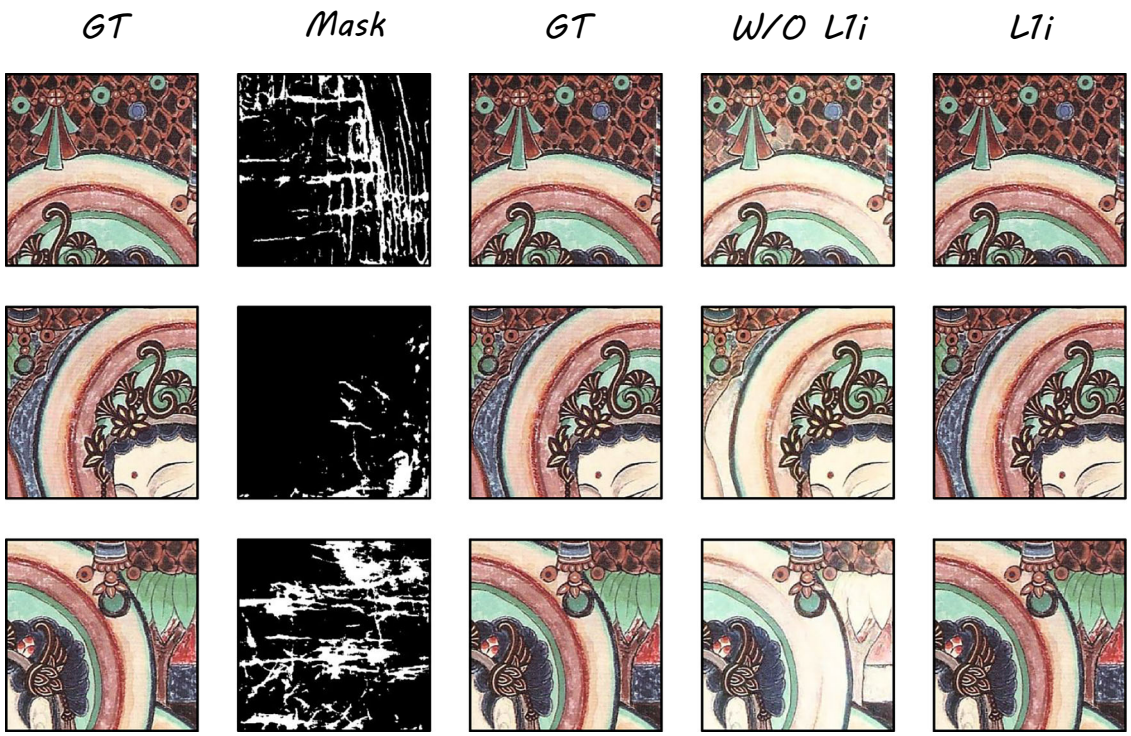

Fig. 14 | Loss Ablation Study. Visual comparison illustrating the effect of the image space L_1 loss. `GT' denotes the ground truth and `Mask' is the damaged input.

Table 4 | Ablation study on the L_1 loss components

Loss Config	PSNR ↑	SSIM ↑	LPIPS ↓
w/o L_{1i}	25.095	0.881	0.024
L_{1i} (Ours)	**26.942**	**0.925**	**0.013**

We compare the performance with and without the L_1 loss on the predicted image (L_{1i}). The test was conducted on masks with a 20%-30% corruption ratio. *Optimal results are displayed in **bold**.

Table 5 | Ablation Study of feature fusion weight

$\lambda_{sketch}/\lambda_{style}$	PSNR ↑	SSIM ↑	LPIPS ↓
0.3/0.7	22.150	0.828	0.017
0.2/0.8	24.964	0.923	0.041
0.1/0.9	**26.942**	**0.925**	**0.013**

Ablation study on the fusion ratio of the sketch and style encoders in DCADif. We evaluate the impact of different weighting coefficients for the sketch (λ_{sketch}) and style (λ_{style}) conditions. *Optimal results are displayed in **bold**.

no variance. This suggests that the main diffusion model is already capable of handling overall color and texture generation.

A minor discrepancy, however, was observed in the SSIM metric. The baseline model without the encoder reached an SSIM of 0.926, slightly higher than the 0.925 achieved with the frozen CLIP. This implies that while the frozen CLIP provides robust structural priors, it may introduce a level of 'rigidity,' potentially limiting the model's flexibility to match local ground truth details perfectly.

Despite this slight drop in SSIM, we incorporate the frozen CLIP encoder in the final design. It acts as a critical structural backbone, ensuring stability and preventing major distortions in large missing regions—benefits that extend beyond what standard metrics can measure. Overall, the results confirm that using a frozen, pre-trained encoder is a robust and effective strategy for ensuring structural fidelity in inpainting.

Effectiveness of Loss Fuction

During the training process, we observed that for diffusion models, relying solely on an L_1 loss to constrain the predicted and ground truth noise can sometimes lead to subtle color discrepancies between the final generated image and the ground truth image, thereby compromising the fidelity of the inpainting. To address this, we introduced an additional L_1 loss term that directly computes the difference between the generated image and the ground truth image, with the aim of enhancing the pixel level alignment capability. To validate the necessity and effectiveness of this design choice, we conducted a key ablation study. The qualitative results are presented in Fig. 14, and the quantitative results are shown in Table 4.

When the L_{1i} loss was removed, a significant decline was observed across all performance metrics. Specifically, the PSNR dropped by a substantial margin of over 1.8 dB, a considerable gap that directly validates the critical role of the L_{1i} loss in color correction and overall fidelity enhancement.

This performance degradation was also evident at the structural and perceptual levels. In the absence of the L_{1i} loss, the SSIM decreased by nearly 5%, indicating that direct image level supervision is crucial for helping the model better reconstruct local structures and ensure the seamless integration of the restored region with its surroundings. The most remarkable change, however, occurred in the LPIPS score, which increased by 84% without the L_{1i} loss. This highlights how theL_{1i}loss effectively aligns the model's optimization objective with the human perceptual space, enabling the generation of more natural-looking images.

Effectiveness of Time Step Ratio

Within TAFF module, the fusion ratio of sketch and style features is not static but changes dynamically over time. To determine the optimal proportion for model performance, we conducted experiments with various feature fusion ratios. The qualitative results are illustrated in Fig. 15, and the quantitative results are presented in Table 5.

The experimental results indicate that the model achieves comprehensively optimal performance when the weight for sketch guidance is set to 0.1 and the weight for style guidance is set to 0.9. However, a deeper analysis of the performance variations under different weight configurations reveals

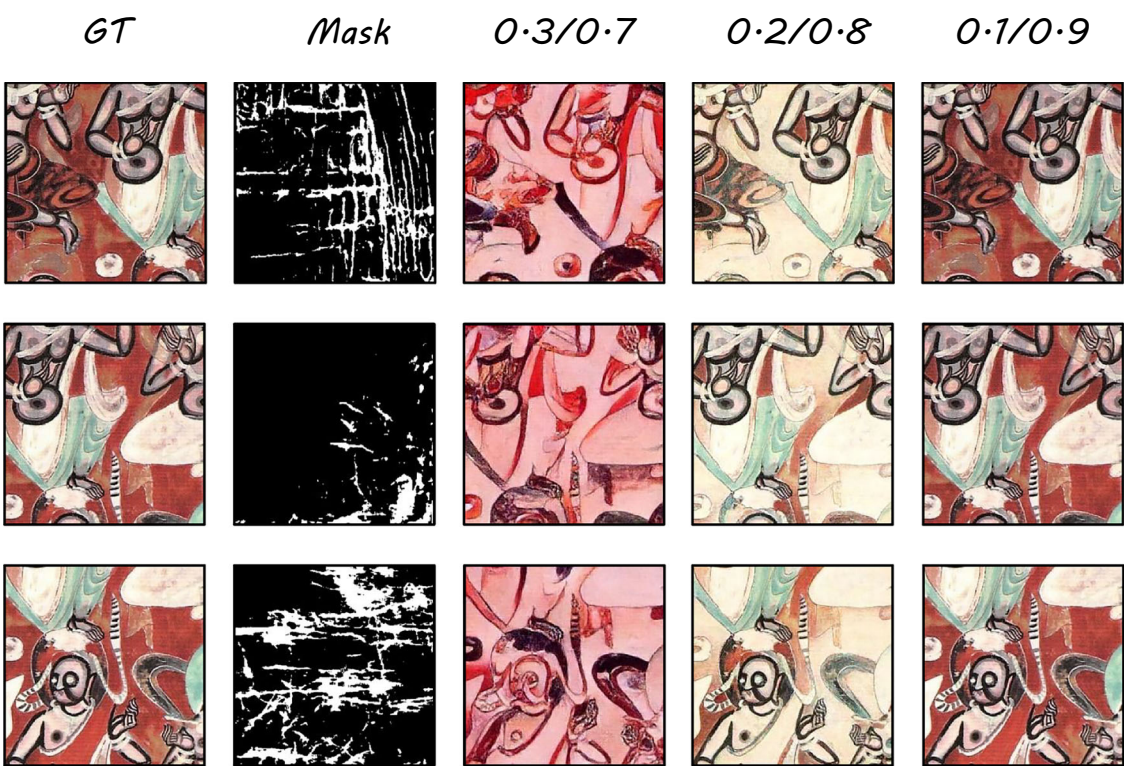

Fig. 15 | Visual ablation study on the fusion ratio for the line art and style conditions in TAFF module. `GT' denotes the ground truth, and `Mask' is the damaged input. The other columns show inpainting results for different [Line Art Weight] / [Style Weight] ratios.

a more profound synergistic and constraining relationship between the two information sources.

The performance exhibits high sensitivity to this fusion ratio. When we increased the sketch weight from 0.1 to 0.2, although the SSIM decreased only negligibly, the PSNR experienced a drastic drop of nearly 2.0 dB, while the LPIPS worsened by more than threefold.

Upon further increasing the sketch weight to 0.3, this trend of performance degradation continued. Both the PSNR and SSIM metrics continued to fall sharply, indicating that an overreliance on the given sketch information impedes the model's ability to learn from the data driven style features and generate natural inpainting content that matches the surrounding context.

In summary, a 0.1/0.9 ratio represents the optimal fusion balance for sketch and style guidance. Style guidance serves as the core driving force for generating highfidelity and photorealistic content, whereas sketch guidance, within the Condition Encoder, should play a subtle, auxiliary corrective role. Overemphasizing external structural constraints severely undermines the model's powerful internally learned generative capabilities. This finding provides a solid experimental basis for the design, validating the effectiveness and rationality of the current weight configuration.

Discussion

We propose an innovative framework for the inpainting of damaged murals, termed DCADif. We employ two encoder modules, the CLIP Sketch Encoder and the SwinStyle Encoder, to learn the deep features of the image in a decoupled and progressive manner. Specifically, within DCADif, we introduce a Time step Adaptive Feature Fusion module. This module deeply couples the denoising process with information injection by dynamically modulating the weights of structural and stylistic features according to the current timestep, thereby prioritizing the establishment of the macro structure in the early denoising stages and the meticulous rendering of micro details in the later stages to achieve a harmonious synthesis of both.

Experimental results demonstrate that DCADif exhibits superior performance in processing murals, particularly in the task of restoring damaged murals, where it shows exceptional capabilities in artistic style preservation and detail inpainting. This validates its effectiveness in the field of cultural heritage preservation and inpainting. Furthermore, the model achieves excellent visual results on a dataset of Chinese paintings, which further substantiates its generalization capability in image inpainting tasks.

Despite the promising results achieved by DCADif in mural inpainting, it is subject to several limitations that offer clear avenues for future research.

First, the model faces challenges when dealing with extremely large, contiguous areas of damage. When critical structural information in a region is completely lost, the line art condition our model relies on becomes unreliable. In such cases, the model may generate content that is visually plausible but historically inaccuratea phenomenon known as hallucination which is a critical concern for the rigorous demands of cultural heritage preservation.

Second, as a diffusion based model, the iterative denoising process of DCADif is computationally intensive. This leads to slower inference speeds compared to single pass architectures like GANs, which could be a practical constraint in applications requiring rapid processing or large batch restoration.

Finally, while our SwinStyle Encoder effectively captures artistic style, its ability to reproduce highly specific material textures could be further improved. For instance, the model may not perfectly distinguish between the unique craquelure patterns on aged plaster and the fibrous texture of ancient silk paintings. Developing more refined restoration algorithms specifically tailored to the material characteristics of cultural artifacts remains an important direction for future work.

Recent studies have shown that frequency aware learning can offer finer control over detail generation across different frequency bands[38]. Incorporating such mechanisms could potentially enhance our model's ability to restore the full spectrum of details in murals from coarse wall textures to delicate brushstrokes thereby further advancing the fidelity of the restoration.

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/DCADif.

Code availability

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

Received: 17 September 2025; Accepted: 14 January 2026;
Published online: 28 January 2026

References

1. He, K. et al. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009 (2022).
2. Yu, J. et al. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4471–4480 (2019).
3. 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*, 10684–10695 (2022).
4. Lugmayr, A. et al. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11461–11471 (2022).
5. Zeng, Y., Fu, Z., Chao, H.-Y. & Zhang, L. CR-Fill: Generative image inpainting with auxiliary contextual reconstruction. In *European Conference on Computer Vision (ECCV)*, 151–167 (Springer, 2022).
6. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022 (2021).
7. Zhu, L. et al. Enlightening low-light images with dynamic guidance for context enrichment. *IEEE Transactions on Circuits and Systems for Video Technology* **32**, 5068–5079 (2022).
8. Barnes, C., Shechtman, E., Finkelstein, A. & Goldman, D. B. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* **28**, 24 (2009).
9. Suvorov, R. et al. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2149–2159 (2022).
10. Zhang, L., Rao, A. & Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847 (2023).
11. Tumanyan, N., Geyer, M., Bagon, S. & Dekel, T. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1921–1930 (2023).
12. Brooks, T., Holynski, A. & Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18392–18402 (2023).
13. Couairon, G., Verbeek, J., Schwenk, H. & Cord, M. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427* (2022).
14. Radford, A. et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PmLR, 2021).
15. Avrahami, O., Lischinski, D. & Fried, O. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18208–18218 (2022).
16. Sharma, D., Dhiman, C. & Kumar, D. Control with style: Style embedding-based variational autoencoder for controlled stylized caption generation framework. *IEEE Transactions on Cognitive and Developmental Systems* (2024).
17. Sharma, D., Dhiman, C. & Kumar, D. Evolution of visual data captioning methods, datasets, and evaluation metrics: A comprehensive survey. *Expert Systems with Applications* **221**, 119773 (2023).
18. Sharma, D., Dhiman, C. & Kumar, D. A review of stylized image captioning techniques, evaluation parameters, and datasets. In *2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST)*, 1–5 (IEEE, 2022).
19. Gatys, L. A., Ecker, A. S. & Bethge, M. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423 (2016).
20. Deng, Y. et al. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11326–11336 (2022).
21. Sharma, D., Dhiman, C. & Kumar, D. Unma-capsumt: unified and multi-head attention-driven caption summarization transformer. *arXiv preprint arXiv:2412.11836* (2024).
22. Wang, Z. et al. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17683–17693 (2022).
23. Rezvani, S., Fateh, M. & Khosravi, H. Abanet: attention boundary-aware network for image segmentation. *Expert Systems* **41**, e13625 (2024).
24. Rezvani, S., Fateh, M., Jalali, Y. & Fateh, A. Fusionlungnet: multi-scale fusion convolution with refinement network for lung ct image segmentation. *Biomedical Signal Processing and Control* **107**, 107858 (2025).
25. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30 (2017).
26. Wang, Z. & Bovik, A. C. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine* **26**, 98–117 (2009).
27. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**, 600–612 (2004).
28. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595 (2018).
29. Yang, K.-F. et al. Adair: Adaptive all-in-one image restoration. In *European Conference on Computer Vision*, 435–452 (Springer, 2022).
30. He, H.-J., Chen, Q.-G., Guo, M.-H., Ouyang, B.-W. & Yan, D.-M. Coupled texture-structure-aware decomposition and guided restoration for image inpainting. *IEEE Transactions on Multimedia* (2023).
31. Li, J., Wang, N., Zhang, L., Du, B. & Tao, D. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7760–7768 (2020).
32. Zhang, L., Zhu, Y., Li, X., Luan, F. & Li, T.-T. Diffusionart: Training-free line art colorization via convolutional adaptive instance normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19523–19532 (2023).
33. Li, W. et al. Mask-aware transformer for large-hole image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11172–11181 (2022).
34. Potlapalli, V., Kulkarni, K. A., Nagaraj, T. & Balasubramanian, V. N. Promptir: Prompting for all-in-one blind image restoration. In *NeurIPS* (2023).
35. Chen, Z. et al. Strdiffusion: Structured momentum diffusion for accelerated image inpainting. In *CVPR* (2024).
36. Song, Y. et al. Score-based generative modeling through stochastic differential equations. In *ICLR* (2021).
37. Lugmayr, A., Danelljan, M., Romero, A., Yu, F. & Timofte, R. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR* (2022).

38. Zhu, L., Zeng, X., Chen, B., Liu, S. & Li, P. Leveraging diffusion knowledge for generative image compression with fractal frequency-aware band learning. *arXiv preprint arXiv:2503.11321* (2025).

Acknowledgements

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

Author contributions

X.P.: Software, Investigation, Validation. C.L.: Preparation, methodology. Q.H.: Conceptualization, software, validation. Z.S.: Editing, Supervision. J.P.: Project administration. M.S.: resources, data curation, Formal analysis.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Qiyao Hu or Manli Sun.