Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Ancient Painting Garment Inpainting via Garment Structure Guidance and Garment Region-Reweighted Flow Matching

Jinjing Yu
Xi'an Polytechnic University
Juan Chai
Xi'an Polytechnic University
Xiyue Zhang
Xi'an Polytechnic University
Rong Fu
Xi'an Polytechnic University
Kaixuan Liu
liukaixuan819@163.com
Xi'an Polytechnic University

Article

Keywords:

Posted Date: August 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10499013/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Ancient Painting Garment Inpainting via Garment Structure Guidance and Garment Region-Reweighted Flow Matching

Jinjing Yu[1], Juan Chai[1], Xiyue Zhang[2], Rong Fu[3], Kaixuan Liu[2] (Corresponding Author)
1. School of Textile Science and Engineering, Xi'an Polytechnic University, Xi'an, China
2. Apparel & Art Design College, Xi'an Polytechnic University, Xi'an, China
3. School of New Media Art, Xi'an Polytechnic University, Xi'an, China
Corresponding Author *E-mail: liukaixuan819@163.com*

Abstract

Garments in ancient paintings often exhibit broken contours and missing drapery details caused by aging. General-purpose inpainting models lack garment-specific structural priors and often overlook small, narrow, and elongated defects. This study proposes a two-stage framework combining Fashion LoRA guidance with Garment Region-Reweighted Flow Matching (GR-FM). First, Fashion LoRA is trained on modern garment line art and structural text to initialize outlines and folds during color restoration. Second, GR-FM normalizes and reweights latent-token regression errors in damaged regions within FLUX.1-Fill, strengthening local supervision. Hard masking is applied during inference to confine modifications. Experiments using four control configurations and three random seeds show that Fashion LoRA mainly improves edge quality in the modern source domain, whereas GR-FM provides more stable cross-domain gains. The complete framework improves pixel, structural, and perceptual metrics on ancient painting datasets and produces digital restoration candidates with traceable spatial boundaries, supporting structure-constrained and auditable cultural heritage restoration.

Ancient paintings constitute important visual records for studying historical dress forms, ritual systems, and changing aesthetic conventions. Collars, cuffs, waistlines, hems, and folds in these paintings collectively define the outer contours and dynamic shapes of the garments. However, deterioration of paper and silk supports, pigment flaking, abrasion of ink lines, and prolonged environmental exposure often disrupt these slender, curved, and spatially continuous lines, substantially reducing the legibility of garment structures[1,2]. Consequently, the restoration of garments in ancient paintings is not merely a problem of texture synthesis but rather a challenge of structural reconstruction under limited historical evidence[3]. Digital restoration must also follow the principles of minimal intervention, process traceability, and explicit identification of generated content[4]. And the resulting images should not be treated as definitive representations of the original historical appearance.

Traditional image-inpainting methods rely primarily on geometric diffusion or exemplar-based patch propagation[5–7]. Although these approaches perform adequately on fine cracks and simple repetitive textures, local propagation cannot reliably preserve

the continuity of structural lines when damage crosses garment contours or interrupts multiple folds[8]. Studies of murals and ancient paintings have likewise shown that improving priority functions or matching strategies alone cannot consistently restore structure-dominated regions without explicitly modeling structural information[9,10].

In recent years, deep generative models have substantially improved semantic reconstruction for complex missing regions. Generative adversarial networks (GANs), diffusion models, and conditional-control frameworks can now address large-area damage[11–13]. LaMa enlarges the receptive field through fast Fourier convolutions, thereby improving long-range feature interactions[14]. More broadly, inpainting architectures have evolved from convolutional and attention-based models toward Transformer and diffusion-based generative frameworks[15]. Contextual attention explicitly aggregates valid features over long distances[16], whereas partial convolution reduces interference from substitute values in masked regions by operating only on valid pixels and renormalizing the responses. Diffusion models further strengthen semantic generation capabilities for large missing regions and enable multimodal control through text, masks, shapes, and related conditions. RePaint generates semantically coherent candidates through iterative denoising[17,18]. Palette unifies inpainting, uncropping, and restoration within an image-to-image generative framework[19]. And ControlNet demonstrates that conditions such as edges, poses, and segmentation maps improve spatial controllability[20]. For localized image editing, the joint control of text and images, protection of the original background, and branch-wise separation of the damaged input from noisy latent variables have also become important design strategies in diffusion-based inpainting[21].

However, general-purpose generative models primarily learn distributions from large-scale natural-image datasets and are optimized for visual realism and semantic plausibility rather than fidelity to historical evidence. A recent review of painting conservation identified the scarcity of high-quality historical training data, the absence of pre-damage references, and distributional differences among artworks as major barriers to machine-learning-assisted virtual restoration[22]. When there is insufficient context in damaged areas, models may generate content that is visually coherent but lacks a clear structural basis. In response to this challenge, previous studies have adopted two-stage frameworks based on structural reconstruction followed by content restoration[23], or introduced global frequency-domain modeling into edge-guided networks to improve the continuity of long-range contours and complex textures. Accordingly, edge-first and structure-guided inpainting paradigms have received increasing attention[24,25]. EdgeConnect first predicts high-frequency edges in the missing region and then uses those edges to guide color and texture synthesis, reducing oversmoothing and edge blurring[26]. For severe, large-area damage in Dunhuang murals, Li et al.[11] proposed a line-sketch-guided progressive restoration method and demonstrated the importance of linear structure. ITS[27], MAT[28], and AOT-GAN[29] subsequently improved long-range structural reasoning through an independent structure-recovery module, dynamic mask-aware attention, and multi-receptive-field contextual transformations, respectively. These works collectively demonstrate that explicit structural modeling helps alleviate edge blurring and connection instability in large-scale damage regions.

In cultural heritage applications, line-sketch-guided mural restoration[11], edge-guided completion of ancient paintings[30–33], and archaeological line-sketch generation from limited reference images[34] all suggest that structural priors help preserve contour continuity, reduce edge blurring, and improve interpretability. Multi-agent prompt-propagation methods have also attempted to integrate evidence from multiple sources within damaged regions[35]. In ancient paintings and murals, where linear contours and color fields jointly define form[36], the direction, spatial position, and overall continuity of lines have a more direct effect on the legibility of a restoration than the photorealism of local textures.

Compared with general objects or background textures, figure garments exhibit stronger object-specific structure. The direction of drapery lines is determined not only by local pixel distributions but also by body pose, garment pattern, component layering, fabric drape, and the direction of applied forces. Garment folds therefore cannot simply be regarded as ordinary edges in natural images, and the connection patterns embody transferable knowledge of garment structure. Modern fashion studies have established datasets and methods for garment detection, landmark localization, human-pose estimation, and semantic segmentation[37,38], providing a foundation for learning statistical regularities in garment structure. Research on the digitization of cultural heritage garments likewise identifies that garment contours, component relationships, and geometric layouts constitute essential forms of evidence for restoring and reconstructing historical garments[3]. Ren et al.[39] demonstrated the value of global frequency-domain modeling for maintaining the continuity of garment edges and textures. In contrast, the present study addresses the scarcity of paired ancient-painting garment data and the insufficient supervision of fine linear damage by combining adaptation on modern garment line drawings with token-level reweighting in latent space.

In addition to structural adaptation, the way a training objective allocates attention across spatial regions also affects the restoration of fine-line damage. Flow matching trains continuous normalizing flows by regressing a conditional vector field along a prescribed probability path, without explicitly simulating the trajectory[40]. Rectified flow enables efficient transport between data distributions along approximately straight paths and has been applied to image generation, image translation, and domain transfer[41]. Standard flow-matching losses, however, generally aggregate errors uniformly across spatial locations or latent tokens. When a mask covers only a few drapery lines or a slender contour, damaged tokens represent only a small proportion of all latent tokens, and their gradients can be diluted by errors from the much larger intact region. Thus, even as the overall loss decreases, the model may fail to learn the structural connectivity of small, line-shaped missing regions adequately.

To address these limitations, this study makes the following contributions:

- A two-stage framework is developed in which Fashion LoRA provides structural initialization and a color-inpainting stage performs task-specific adaptation, thereby guiding the recovery of garment contours and folds.
- Garment Region-Reweighted Flow Matching (GR-FM) is introduced to normalize and reweight velocity-field errors for latent tokens in damaged regions, alleviating

insufficient supervision for small, narrow, elongated defects without adding an inference-time module.

- A complete four-configuration controlled experiment with three random seeds and paired statistical tests is conducted to distinguish the effects of structural initialization, region-focused supervision, and their conditional interaction, while also examining the limits of target-domain applicability.
- Hard-mask compositing restricts all modifications to manually annotated regions, providing restoration candidates with explicit and reviewable spatial boundaries.

Methods

Overall methodology framework

The proposed method comprises three components: A, preparation of ancient painting data. B, adaptation of garment structure guidance to the color defect restoration task. and C, local inpainting of garments in ancient paintings. Step B includes two consecutive training phases, as shown in Figure 1.

First, Florence-2 is used to assist in localizing individual figures. After manual verification, the detected regions are converted into single-person semantic crops and 512×512-pixel garment-detail patches. The ancient-painting data are used only for target-domain evaluation and validation on genuine damage and do not participate in model training.

Next, two-stage training is performed on modern garment images. In the first stage, the garment images are converted into line drawings that emphasize contours and folds, and Fashion LoRA is trained using these drawings together with structural text. In the second stage, color inpainting is performed with FLUX.1-Fill, while GR-FM strengthens supervision for latent tokens in damaged regions. Fashion LoRA affects parameter initialization, whereas GR-FM modifies the training objective.

Finally, the LoRA adapter obtained from the second stage is applied to localized inpainting of ancient-painting garments, and hard-mask compositing retains generated pixels only within the annotated regions.

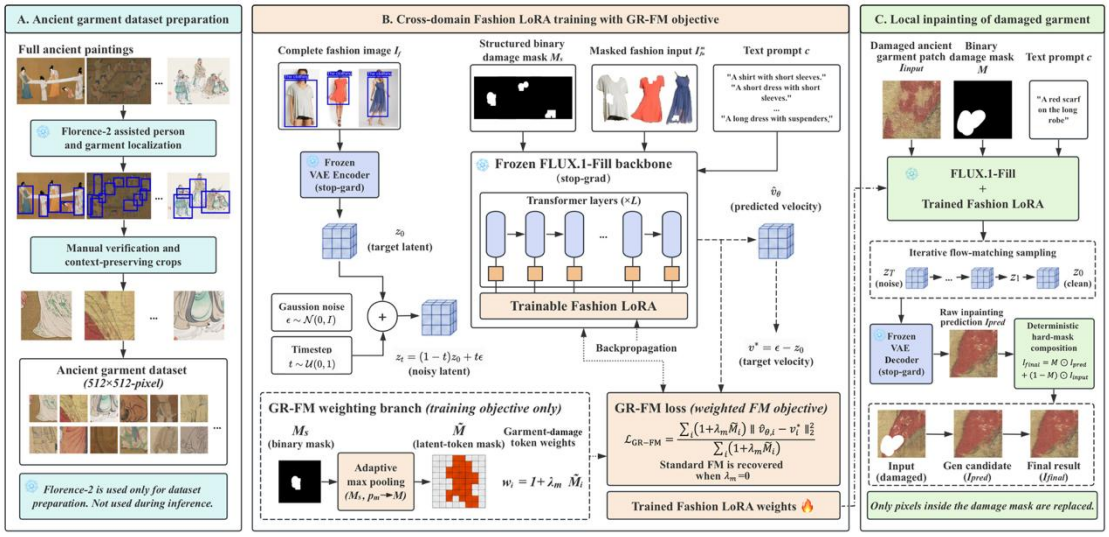


Fig. 1 | Overall framework for inpainting damaged garments in ancient paintings.

Construction of ancient painting garment dataset

The ancient-painting dataset covers five historical periods: the Tang, Song, Yuan, Ming, and Qing dynasties. The paintings vary substantially in the number of figures and the complexity of their backgrounds. Directly resizing an entire painting would weaken garment details and introduce irrelevant background content.

Florence-2 is first used to obtain candidate bounding boxes for figures[42], as shown in Figure 2. Each candidate box is manually reviewed to ensure that it fully covers the principal garment and is adjusted according to pose, visible extent, and damage location. A small amount of surrounding context is retained to avoid truncating long sleeves, hems, or folds that extend across regions. In this study, Florence-2 is used exclusively as an auxiliary localization tool during ancient-painting data preparation, specifically to:

- assist in identifying figure locations in the original ancient paintings;
- determine the principal spatial extent of each figure's garments;
- provide candidate crop regions for standardized ancient-painting garment patches.

Based on the single-person semantic crops, an additional 512×512-pixel detail patches are extracted around collars, cuffs, waistlines, hems, and major folds. Each patch retains records of the source image, figure information, and historical period. The single-person semantic crops and detail patches form a hierarchical dataset and are not counted as independent samples.

After data preparation, the ancient-painting images are excluded from parameter training for both Fashion LoRA and GR-FM. They are used for target-domain evaluation, trend analysis, and validation on genuinely damaged cases.

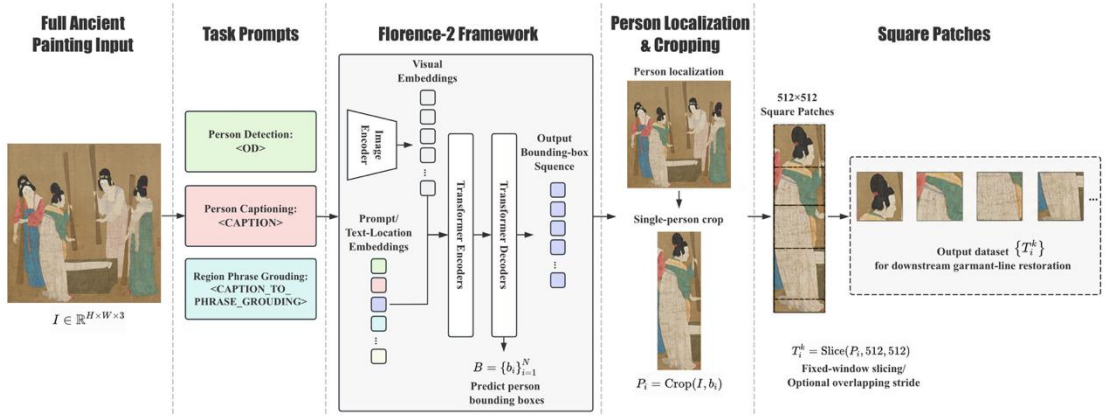


Fig. 2 | Construction of the semantic restoration units.

Garment-structure guidance and Fashion LoRA

Construction of modern garment line drawings and color-inpainting samples

Genuinely damaged ancient paintings generally lack pre-damage references, making it difficult to construct a large-scale paired training dataset. Modern garment images are therefore used to learn the structural characteristics of garment contours and folds.

Dresses and other long or loose-fitting garments with complete contours, limited occlusion, and high image quality are selected from the modern-garment training set. Figures, exposed body parts, and backgrounds are removed before line drawings are generated. After the training-validation split is completed, color-inpainting samples are created using the following structured masks:

- thin, elongated missing regions distributed along garment contours;
- curved strip-shaped missing regions that interrupt folds;
- localized missing regions covering cuffs, waistlines, or hems;
- compact irregular missing regions crossing multiple structural boundaries.

The line drawings retain the outer contour, cuffs, hems, waistlines, and major folds, thereby suppressing appearance-related factors such as photographic lighting, material highlights, and background textures. The structured masks are used only for color reconstruction in the second stage.

Each line drawing is paired with structural text to form a Fashion LoRA training sample. The text includes the trigger token "glineprior" and descriptions of contours, folds, and drape, enabling LoRA to learn the correspondence between the trigger token and garment line structure.

Fashion LoRA parameter tuning

FLUX.1-Fill provides general image-generation and localized inpainting capabilities but is not specifically adapted to fine garment lines or continuous folds. This study employs LoRA for parameter-efficient adaptation: the backbone of the base model remains frozen, and only the low-rank parameters are updated.

For the base weights W, the adapted weight is expressed as:

$$W' = W + \frac{\alpha}{r} BA \tag{1}$$

where A and B are trainable low-rank matrices, r is the low-rank, and α is the scaling factor.

The adapter obtained in the first stage is termed Fashion LoRA. It encodes source-domain structural biases associated with garment contours, hems, cuffs, and folds. In the second stage, random LoRA initialization and Fashion LoRA initialization are treated as separate conditions to distinguish the independent effects of structural initialization and regional reweighting.

GR-FM: Garment Region-Reweighted Flow-Matching Loss

Standard flow matching

During color reconstruction of damaged garments, the complete color target is first encoded into latent space by a frozen variational autoencoder (VAE). Gaussian noise and a time step are then sampled to construct the intermediate latent variable:

$$z_t = (1-t)z_0 + t\grave{o} \tag{2}$$

For this linear probability path, the target velocity field is defined as:

$$v^* = \frac{\mathrm{d}z_t}{\mathrm{d}t} = \grave{o} - z_0 \tag{3}$$

Given the intermediate latent variable z_t, time step t, and conditioning information c, the model predicts the corresponding velocity field:

$$\hat{v}_\theta = v_\theta\left(z_t, t, c\right) \tag{4}$$

where θ denotes the trainable LoRA parameters. The conditioning information c includes the damaged color garment image, the binary damage mask, and the text condition. The FLUX.1-Fill backbone and VAE parameters remain frozen, and only the low-rank LoRA parameters are updated.

Assume that the latent space contains N tokens. For the i-th token, the model produces a predicted velocity and a target velocity. The standard flow-matching loss assigns equal weight to the velocity-field regression error of every latent token:

$$\mathrm{L}_{\mathrm{FM}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{v}_{\theta,i} - v_i^*\right\|_2^2 \tag{5}$$

The model predicts the velocity field from the intermediate latent variable, time step, damaged color input, binary mask, and text condition. The discrepancy between the predicted and target velocities is used to update the LoRA parameters.

Standard flow matching weights all latent tokens equally. For narrow folds or contour damage, damaged tokens account for only a small fraction of the latent space. Consequently, intact tokens may dominate the optimization and weaken supervision related to structural recovery.

GR-FM loss

To enhance training supervision for small-area missing regions, this study proposes GR-FM, a reweighted flow matching loss for garment missing regions, as shown in Figure 3. This method does not alter the velocity field prediction process of FLUX.1-Fill but instead uses a known binary mask to normalize and reweight the regression errors of latent variable tokens.

First, adaptive max pooling maps the binary mask from pixel space onto the latent-token grid, producing a token-level mask. Any latent token that covers at least one damaged pixel is labeled as a damaged-region token, reducing the risk that narrow masks disappear during downsampling.

The weight of the i-th token is defined as:

$$w_i = 1 + \lambda_m M_i \tag{6}$$

where $\lambda_m \geq 0$ is the reweighting coefficient for damaged tokens. The weight of each intact token remains 1, whereas the weight of each damaged token is $1 + \lambda_m$. The GR-FM loss is then expressed as:

$$\mathrm{L}_{\text{GR-FM}} = \frac{\sum_i w_i \left\|\hat{v}_{\theta,i} - v_i^*\right\|_2^2}{\sum_i w_i} \tag{7}$$

The denominator normalizes the loss by the sum of token weights, controlling the overall loss scale across different λ_m settings and preventing apparent performance changes from arising solely from a uniform increase in the loss magnitude.

When $\lambda_m = 0$, GR-FM reduces to standard flow matching. To identify the respective effects of Fashion LoRA initialization and GR-FM, we construct four controlled configurations:

- No Fashion LoRA + Standard FM: the second-stage LoRA is randomly initialized, with $\lambda_m = 0$;
- Fashion LoRA + Standard FM: the first-stage Fashion LoRA is loaded, with $\lambda_m = 0$;
- No Fashion LoRA + GR-FM: the second-stage LoRA is randomly initialized, with $\lambda_m = 20$;
- Fashion LoRA + GR-FM: the first-stage Fashion LoRA is loaded, with $\lambda_m = 20$.

All four configurations use the same FLUX.1-Fill backbone, color-inpainting training data, LoRA architecture, learning rate, number of training steps, and inference settings. The two experimental factors are Fashion LoRA initialization and the GR-FM training objective; all other conditions are held constant.

GR-FM is used only during second-stage color-inpainting training. It is not applied during first-stage line-drawing-text structural adaptation and does not introduce any additional inference module. After training, its effect is encoded in the final LoRA parameters.

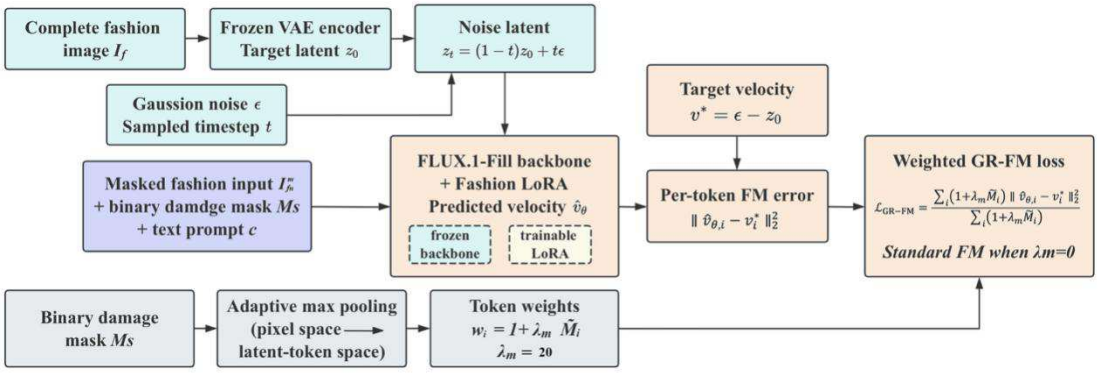


Fig. 3 | GR-FM garment region-reweighted flow-matching loss.

Local inpainting of damaged garment in ancient paintings

After training, the LoRA adapter further optimized during the color-inpainting stage is loaded into FLUX.1-Fill. The inference inputs consist of a damaged color patch from an ancient painting, its corresponding binary damage mask, and the text condition.

When contour or fold cues remain on both sides of a damaged region, the model tends to generate connecting structures with consistent directions. However, the unobserved content cannot be uniquely determined. Therefore, this study defines the output as a candidate for digital restoration rather than a deterministic reconstruction of the original historical appearance.

The final image is obtained through hard-mask compositing:

$$I_{\text{final}} = M \odot I_{\text{pred}} + (1 - M) \odot I_{\text{input}} \tag{8}$$

where I_{pred} is the generated candidate image, I_{input} is the original damaged input, and M is the binary region in which modifications are permitted.

The generated content can be pasted back into the original painting using the initial crop coordinates. The mask, model version, inference parameters, and candidate result should be stored together to support subsequent review.

Experimental data and implementation details

Data split

The data consist of a modern-garment source domain and an ancient-painting target domain, as shown in Table 1.

After figure localization and manual verification, the ancient-painting data yielded 807 single-person semantic crops, from which 2,722 garment-detail patches of 512×512 pixels were further extracted. The two types of crops form a hierarchical relationship and are not added together as independent samples. The ancient-painting data are not used for model training. They are reserved for quantitative target-domain evaluation and validation on genuinely damaged cases.

The modern-garment dataset initially contained 1,000 candidate images. After garment extraction, background removal, and manual quality screening, 850 valid images were retained. The data were split into training and validation sets at the level of the original garment image. Each original image and all of its derived damaged samples were restricted to a single subset to prevent leakage between samples originating from the same source image.

From the modern-garment training subset, 500 images with complete contours, clear folds, and high-quality edges were selected. After edge extraction, these images produced 500 garment line-drawing-structural-text pairs for Fashion LoRA adaptation to line structure. This subset was drawn from the modern-garment training data and does not constitute an additional independent data source.

Color-inpainting samples were generated only after the training-validation split. Structured masks were sampled at multiple locations along garment contours, folds, cuffs, waistlines, and hems, producing 1,500 color-inpainting training pairs and 200 fixed validation pairs. Each pair contains a complete color target, a damaged color input, a binary damage mask, and a text condition, as illustrated in Figure 4. Validation pairs were not used for parameter updates.

The ancient-painting target-domain evaluation set contains 50 garment-detail samples with complete reference images. Each sample includes the complete reference, a manually constructed binary mask, and the corresponding damaged input.

The genuinely damaged ancient-painting set contains 20 garment samples without complete references and covers pigment loss, abrasion of paper or silk supports, low-contrast line breaks, and irregular surface loss.

Table 1 | Composition and purpose of the experimental datasets.

Data Domain	Data Subset	Number	Data Composition	Primary Purpose

Modern-garment source domain	Fashion LoRA line-drawing training set	500	Garment line drawings generated by the edge-extraction pipeline and structural text	Learning structural biases in garment contours and folds
Modern-garment source domain	Color-inpainting training set	1,500	Complete color target, damaged color input, binary mask, and text	Training Standard FM and GR-FM and conducting parameter-sensitivity analysis
Modern-garment source domain	Color-inpainting validation set	200	Same composition as the training set, with fixed damage masks	Component ablation, parameter-sensitivity analysis, and source-domain evaluation
Ancient-painting target domain	Ancient-painting target-domain evaluation set	50	Complete reference, manually constructed mask, and damaged input	Baseline comparison and quantitative target-domain evaluation
Ancient-painting target domain	Genuinely damaged ancient-painting set	20	Genuinely damaged garment regions and manually annotated masks	Reference-free qualitative validation

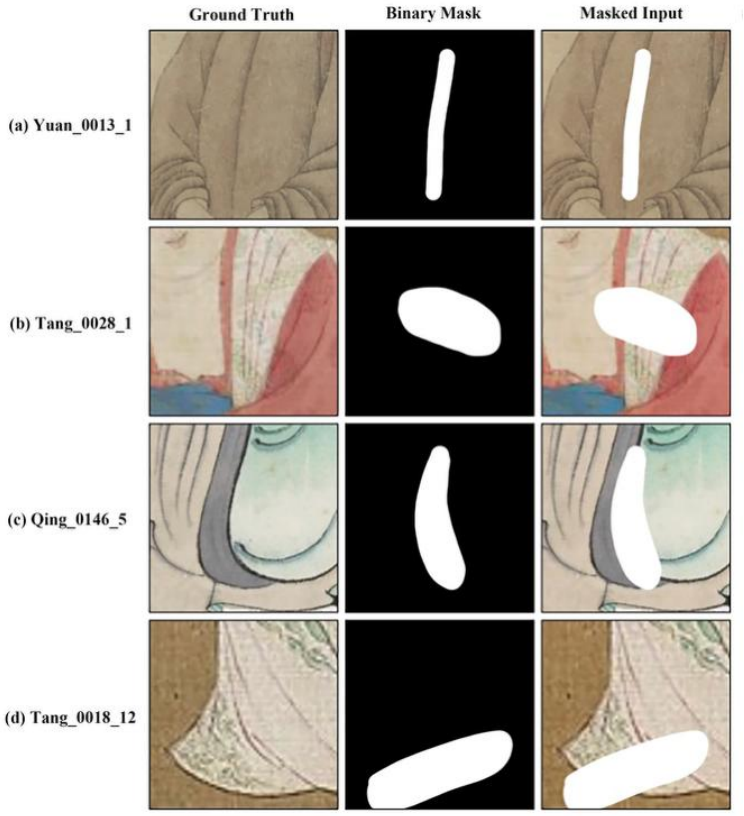


Fig. 4 | Data and task design for inpainting damaged garments in ancient paintings.

Training settings

The experiments were implemented in Python and PyTorch on a workstation equipped with an Intel Xeon W7-3445 processor, 128 GB of memory, and three NVIDIA RTX A6000 GPUs with 48 GB of memory each.

The Fashion LoRA line-structure adaptation utilized 500 pairs of modern garment line drawings and structural text descriptions. The line drawings were generated from

modern garment images through an edge-extraction pipeline, and the text described garment contours, folds, and draping structure. During training, the FLUX.1-Fill backbone was frozen, and only the low-rank LoRA parameters were updated.

The second stage used 1,500 modern-garment training pairs and formed four controlled configurations by varying Fashion LoRA initialization and the training objective. All configurations shared the FLUX.1-Fill backbone, training data, and optimization settings: a learning rate of 5×10^{-5}, a batch size of 1, four gradient-accumulation steps, a LoRA rank and α of 8, a dropout of 0.05, and 500 training steps. Each configuration was trained with three random seeds (42, 3407, and 2026), producing 12 models in total, and the weights from step 500 were used for all evaluations.

To examine the effect of the damaged-token reweighting coefficient, Fashion LoRA initialization was held fixed, and the following values were evaluated on the modern-garment validation set:

$$\lambda_m \in \{0,5,10,20,30,50\} \tag{9}$$

The $\lambda_m = 0$ and $\lambda_m = 20$ settings reused the Fashion + Standard FM and Fashion + GR-FM models, respectively, from the four controlled configurations. Separate models for $\lambda_m = 5$, 10, 30, and 50 were trained with each of the three random seeds (42, 3407, and 2026). The six coefficients therefore corresponded to 18 models. Except for λ_m, the training data, Fashion LoRA initialization weights, LoRA architecture, optimization parameters, number of training steps, and inference settings were identical.

Inference used a LoRA scale of 1.0, 25 sampling steps, and a guidance scale of 3.5. The same inference seed was used for each validation sample across all settings, and the same hard-mask compositing procedure was applied.

Parameter selection jointly considered BF1-2, Masked PSNR, and Masked LPIPS. Once the primary edge metric approached saturation, a moderate reweighting coefficient that provided clear pixel-level and perceptual gains was selected, rather than choosing the local optimum of a single metric.

Evaluation and statistical analysis

Restoration quality was evaluated using Masked PSNR, Masked SSIM[43], Masked LPIPS[44], Boundary F1 with a two-pixel tolerance (BF1-2), and Boundary Edge Recall (BER). Evaluation was restricted to the masked region and its boundary neighborhood, as summarized in Table 2.

Masked PSNR, Masked SSIM, and Masked LPIPS measure pixel fidelity, structural similarity, and perceptual difference within the masked region, respectively.

BF1-2 measures the overall accuracy of garment-edge restoration. Canny edges are extracted from the reference and restored images and matched within an evaluation region formed by dilating the damage mask outward by two pixels, with a positional tolerance of two pixels. Edge precision and recall are defined as:

$$P_b = \frac{N_{\text{match}}^{\text{pred}}}{N_{\text{edge}}^{\text{pred}}}, R_b = \frac{N_{\text{match}}^{\text{ref}}}{N_{\text{edge}}^{\text{ref}}} \tag{10}$$

$$\text{BF1-2} = \frac{2P_bR_b}{P_b + R_b} \tag{11}$$

A predicted edge and a reference edge are considered a valid match when their minimum Euclidean distance does not exceed two pixels. BF1-2 penalizes both spurious generated edges and omitted reference edges; higher values indicate better performance.

BER evaluates whether reference garment edges are continued where they enter or leave the damaged region. It is effectively a recall measure for boundary-related edges and does not penalize spurious generated edges; it should therefore be interpreted together with BF1-2.

$$\mathrm{BER} = \frac{N_{\text{continued}}^{\text{boundary}}}{N_{\text{reference}}^{\text{boundary}}} \tag{12}$$

The denominator is the number of valid reference edges within the boundary band, and the numerator is the number of those edges that are continued in the restored result. If a sample contains no valid reference edge within the boundary band, its BER is treated as missing and excluded from the BER mean and significance tests.

Table 2 | Restoration-quality metrics and optimization directions.

Metric	Evaluation Content	Optimization Direction
Masked PSNR	Pixel fidelity in the missing region	Higher is better
Masked SSIM	Local Structural Similarity in Defective Regions	Higher is better
Masked LPIPS	Perceptual Feature Differences in the Missing Region	Lower is better
BF1-2	Overall accuracy of garment edge restoration	Higher is better
BER	Ability to extend lines at the edges of defects	Higher is better

Results for the three random seeds are reported as the mean of the three seeds and the sample standard deviation, using the mean for each seed on the full dataset as the base unit.

For the four controlled configurations, predictions from the three training seeds were first averaged at the image level. Four controlled comparisons were then assessed using two-sided paired Wilcoxon signed-rank tests, with Holm correction applied to the five metrics within each dataset and comparison[45].

The interaction effect was defined at the image level as a difference in differences:

$$\Delta int = \left(Y_{F,G} - Y_{F,S}\right) - \left(Y_{N,G} - Y_{N,S}\right) \tag{13}$$

where F and N denote the use of Fashion LoRA initialization and the absence thereof, respectively, while G and S denote GR-FM and Standard FM. The interaction term was evaluated using a two-sided Wilcoxon signed-rank test, with Holm correction across the five metrics within each dataset and a significance level of 0.05.

In the λ_m sensitivity plots, each point and error bar represent the dataset-level mean and sample standard deviation, respectively, across three independently trained models. These values were used to assess the overall trend and identify a stable parameter range.

Results

Overall performance in the ancient-painting target domain

To evaluate the overall performance of the complete two-stage model, LaMa, SDXL Inpainting[46], BrushNet-SDXL[47], and FLUX-Fill Base were selected as comparison methods. All methods used the same damaged inputs, masks, resolution, and hard-mask compositing procedure. The results are reported in Table 3.

Table 3 | Quantitative comparison of five methods on the standard ancient-painting evaluation set. Values are reported as the mean ± sample standard deviation across 50 test images. For Fashion LoRA + GR-FM, predictions from the three training seeds were first averaged at the image level.

Method	Masked PSNR ↑	Masked SSIM ↑	Masked LPIPS ↓	BF1-2 ↑	BER ↑
LaMa	8.738 ± 6.020	0.403 ± 0.167	0.785 ± 0.221	0.270 ± 0.172	**0.955 ± 0.115**
SDXL Inpainting	19.092 ± 5.037	0.461 ± 0.152	0.545 ± 0.148	0.417 ± 0.206	**0.920 ± 0.105**
BrushNet-SDXL	18.270 ± 3.380	0.363 ± 0.143	0.505 ± 0.097	0.506 ± 0.205	**0.921 ± 0.083**
FLUX-Fill Base	18.903 ± 7.761	0.472 ± 0.131	0.391 ± 0.113	0.404 ± 0.263	**0.942 ± 0.077**
Fashion LoRA + GR-FM	**22.779 ± 3.877**	**0.535 ± 0.151**	**0.379 ± 0.105**	**0.549 ± 0.206**	0.920 ± 0.092

As shown in Table 3, Fashion LoRA + GR-FM achieved the best mean results for Masked PSNR, Masked SSIM, Masked LPIPS, and BF1-2, with values of 22.779 dB, 0.535, 0.379, and 0.549, respectively; its BER was 0.920. Relative to FLUX-Fill Base, which uses the same backbone, Masked PSNR increased by 3.876 dB, Masked SSIM increased by 0.063, Masked LPIPS decreased by 0.012, and BF1-2 increased by 0.145.

Compared with LaMa, SDXL Inpainting, and BrushNet-SDXL, the complete system achieved better mean values for Masked PSNR, Masked SSIM, Masked LPIPS, and BF1-2, indicating strong overall competitiveness. These findings show that structural initialization learned from the modern-garment source domain, together with GR-FM, performs competitively on the current evaluation set. Table 3, however, neither separates the contributions of Fashion LoRA and GR-FM nor establishes statistical significance across methods. Their respective mechanisms are examined through the four controlled configurations.

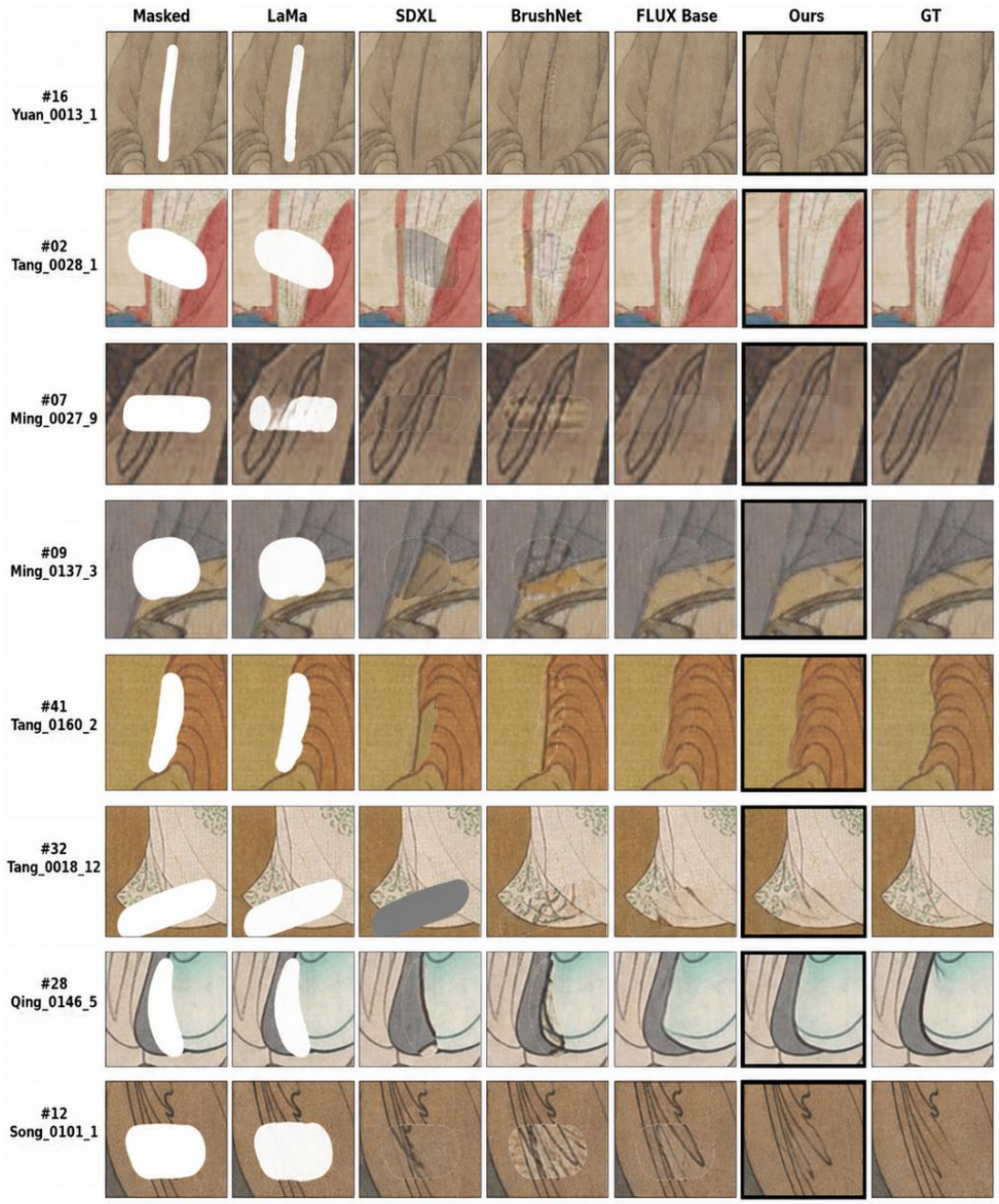


Fig. 5 | Qualitative comparison of representative garment-inpainting results from ancient paintings.

Figure 5 reveals distinct generation tendencies among the baselines. LaMa tends to produce blurred propagation within narrow damaged regions. SDXL Inpainting can complete local content, but some generated drapery lines are inconsistent with the direction of the surviving structure. BrushNet-SDXL may produce duplicated edges, whereas FLUX-Fill Base remains unstable when connecting long-range drapery lines. In most cases, the proposed method more effectively continues contours and folds across the damaged region, although multiple plausible solutions remain when internal evidence is insufficient.

Effects of garment-structure guidance and GR-FM

Qualitative edge-level comparison

To distinguish the effects of Fashion LoRA structural initialization and the GR-FM training objective, four controlled configurations were evaluated with three independently trained random seeds under identical data, training, and inference conditions.

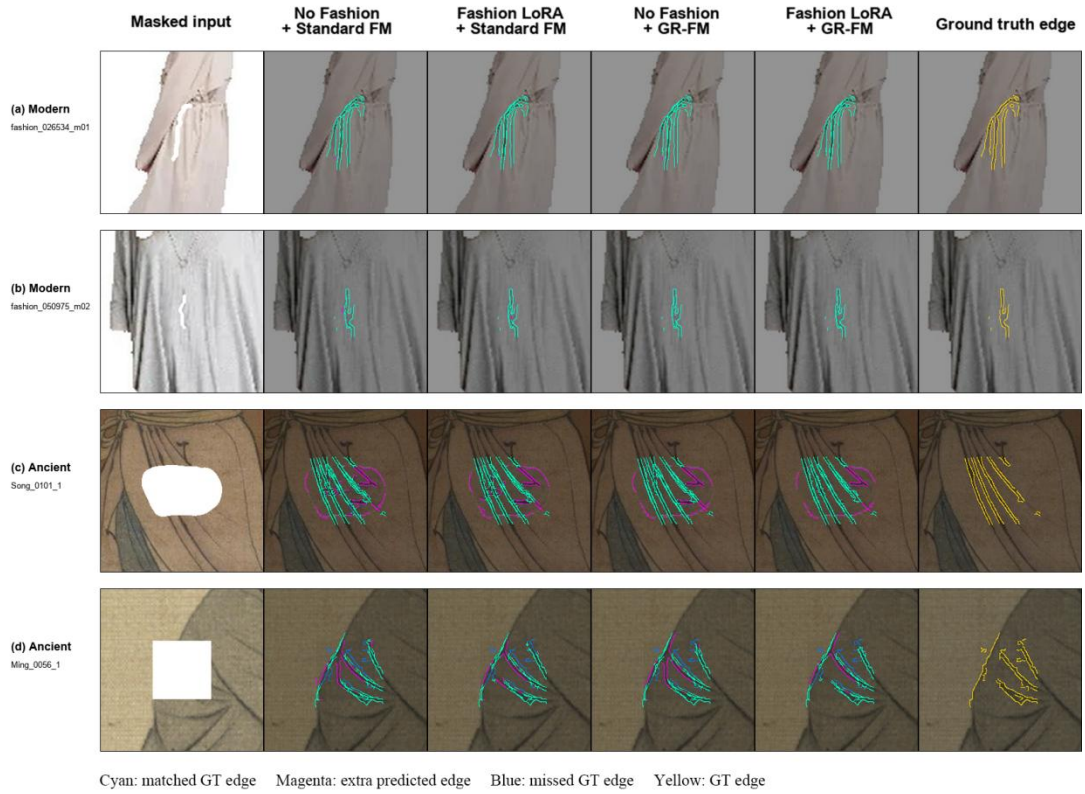


Fig. 6 | Edge-level visualization of the four controlled ablation configurations on modern garments and ancient-painting garments.

From left to right, the columns show the damaged input, No Fashion + Standard FM, Fashion + Standard FM, No Fashion + GR-FM, Fashion + GR-FM, and the reference edges. Cyan indicates predicted edges matched to the reference, magenta indicates additional predicted edges, blue indicates omitted reference edges, and yellow in the rightmost column indicates ground-truth reference edges. All results use training seed 42.

As shown in Fig. 6, matched edges dominate in the modern-garment samples, whereas the ancient-painting target domain contains more spurious and omitted edges, reflecting a clear domain gap. Visual differences among the four configurations are concentrated mainly in local fine lines and contours. Therefore, quantitative metrics and paired statistical tests are required for reliable interpretation.

Quantitative ablation of the controlled comparisons

Table 4 shows that Fashion LoRA primarily improves edge structure in the modern-garment source domain. All p values reported below are Holm-adjusted. Under Standard FM, BF1-2 and BER improved significantly ($p = 5.61\times10^{-7}$ and 7.73×10^{-4}, respectively), and under GR-FM, both metrics again improved ($p = 2.54\times10^{-5}$ and 8.13×10^{-3},

respectively). The remaining pixel-level and perceptual metrics did not show consistent improvement.

In the ancient-painting target domain, the effect of Fashion LoRA was metric-dependent. Under Standard FM, Masked LPIPS increased, and BF1-2 decreased (p = 0.036 and 0.046, respectively). Under GR-FM, Masked SSIM increased (p=0.037), but BF1-2 decreased (p = 0.003). These results indicate that source-domain structural biases do not translate unconditionally into better drapery-line restoration in ancient paintings.

GR-FM demonstrated more consistent task adaptability. Without Fashion LoRA, GR-FM increased BF1-2 on modern garments (p = 0.004) and reduced Masked LPIPS on ancient paintings (p = 0.026). With Fashion LoRA initialization, GR-FM improved Masked PSNR, Masked SSIM, and Masked LPIPS on modern garments ($p = 2.73\times10^{-6}$, 1.24×10^{-10}, and 1.66×10^{-16}, respectively) and improved the same three metrics in the ancient-painting target domain ($p = 1.81\times10^{-7}$, 1.81×10^{-7} and 4.21×10^{-7}, respectively). Regional reweighting thereby enhances supervision for missing tokens and mitigates initialization bias.

Table 4 | Four controlled configurations evaluating Fashion LoRA structural initialization and GR-FM. Values represent the mean and sample standard deviation of dataset-level means across three independently trained models.

Dataset	Configuration	Masked PSNR ↑	Masked SSIM ↑	Masked LPIPS ↓	BF1-2 ↑	BER ↑
Modern	**No Fashion + Standard**	**17.231 ± 0.059**	0.598 ± 0.003	0.113 ± 0.001	0.863 ± 0.002	0.885 ± 0.003
Modern	**Fashion + Standard**	16.952 ± 0.082	0.603 ± 0.004	0.113 ± 0.002	0.885 ± 0.001	**0.909 ± 0.001**
Modern	**No Fashion + GR-FM**	17.074 ± 0.073	0.600 ± 0.004	0.111 ± 0.002	0.875 ± <0.001	**0.897 ± 0.002**
Modern	**Fashion + GR-FM**	17.160 ± 0.211	**0.616 ± 0.007**	**0.104 ± 0.003**	**0.890 ± 0.002**	**0.909 ± 0.002**
Ancient	**No Fashion + Standard**	22.757 ± 0.134	0.531 ± 0.003	0.378 ± 0.003	0.553 ± 0.001	**0.923 ± 0.001**
Ancient	**Fashion + Standard**	22.522 ± 0.028	0.526 ± 0.001	0.386 ± 0.001	0.547 ± 0.002	0.919 ± <0.001
Ancient	**No Fashion + GR-FM**	**22.805 ± 0.278**	0.532 ± 0.007	**0.374 ± 0.006**	**0.558 ± 0.006**	0.920 ± 0.002
Ancient	**Fashion + GR-FM**	22.779 ± 0.017	**0.535 ± 0.001**	0.379 ± 0.003	0.549 ± 0.005	0.920 ± 0.003

Conditional interaction between Fashion LoRA and GR-FM

The Holm-corrected p-values for the interaction tests indicated statistically significant interactions for Masked PSNR, Masked SSIM, and Masked LPIPS on modern garments ($p = 3.21\times10^{-8}$, 6.64×10^{-6} and 5.65×10^{-6}, respectively). Significant interactions were also observed for Masked PSNR and Masked SSIM in the ancient-painting target domain ($p = 9.88\times10^{-6}$and 5.47×10^{-5}, respectively). BF1-2 and BER were not significant for modern garments (both p = 0.216), and Masked LPIPS, BF1-2, and BER were not significant for the ancient-painting target domain (all p = 0.740).

Therefore, the two mechanisms are not simply additive. Fashion LoRA provides source-domain initialization of garment structure, whereas GR-FM improves adaptation to the color-inpainting task by redistributing training supervision. The target-domain benefits of Fashion LoRA are bounded and metric-dependent, while GR-FM is the more stable core mechanism in the present experiments.

The GR-FM configuration without Fashion LoRA achieved better Masked PSNR, Masked LPIPS, and BF1-2 in the ancient-painting target domain, whereas the complete two-stage configuration performed better in Masked SSIM and in the combined structural and perceptual metrics of the modern-garment source domain. Fashion LoRA + GR-FM is used as the complete two-stage configuration for baseline comparison. It was not selected retrospectively on the basis of the ancient-painting evaluation set. This result further indicates that GR-FM is the primary stable source of target-domain performance, whereas the structural initialization supplied by Fashion LoRA remains domain and metric dependent.

Sensitivity analysis of GR-FM reweighting coefficient

To examine the influence of damaged-token reweighting strength, Fashion LoRA initialization was fixed, and λ_m was evaluated over {0, 5, 10, 20, 30, 50}. The $\lambda_m = 0$ and $\lambda_m = 20$ settings reused the Fashion + Standard FM and Fashion + GR-FM models, respectively, from the four controlled configurations. Models for the remaining coefficients were independently trained with three random seeds using identical data, initialization weights, and training settings.

As shown in Fig. 7, Masked PSNR and Masked SSIM generally increased as λm increased, whereas Masked LPIPS gradually decreased, indicating that stronger damaged-token supervision improves local reconstruction quality. BF1-2 increased from 0.885 ± 0.001 at $\lambda_m = 0$ to 0.890 ± 0.002 at $\lambda_m = 10$ and then changed only marginally. Within the range $\lambda_m = 10$–30, the largest difference in mean BF1-2 was only 0.000131, indicating that edge-restoration performance had essentially saturated.

When λ_m increased further from 20 to 50, Masked PSNR increased by 0.101 dB, Masked SSIM increased by 0.0037, and Masked LPIPS decreased by 0.0022, whereas BF1-2 and BER increased by only 0.0008 and 0.0004, respectively. All of these changes were smaller than the corresponding standard deviations across training seeds. Thus, stronger reweighting continued to provide gradual gains, but with limited marginal returns. Then selected $\lambda_m = 20$ as a moderate reweighting setting that balances edge restoration, pixel fidelity, perceptual quality, and reweighting strength. This choice was not based on the highest value of any single metric.

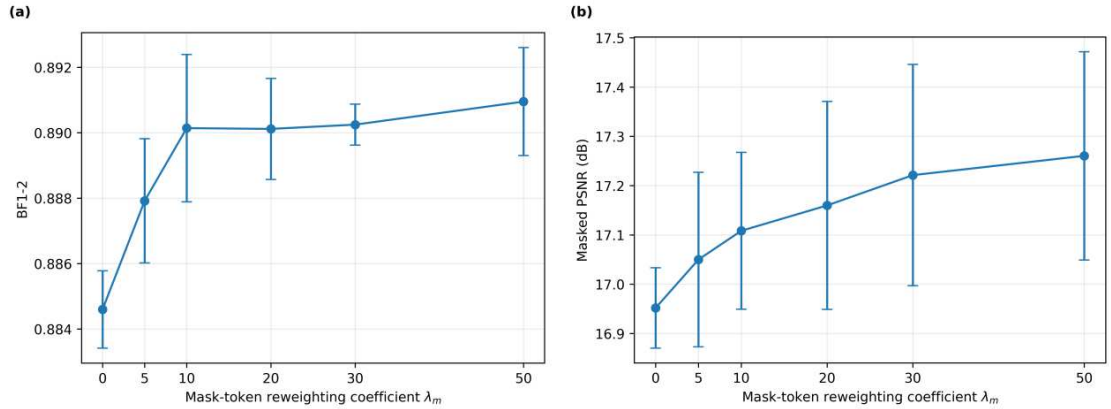


Fig. 7 | Sensitivity analysis of the GR-FM damaged-token reweighting coefficient.

Accordingly, $\lambda_m = 20$ was selected near the center of the stable plateau to balance structural restoration, perceptual quality, and parameter stability, rather than to maximize a single metric locally.

Validation on genuinely damaged ancient paintings

Because genuinely damaged samples do not have complete reference images, they were evaluated qualitatively only. Figure 8 presents six representative cases.

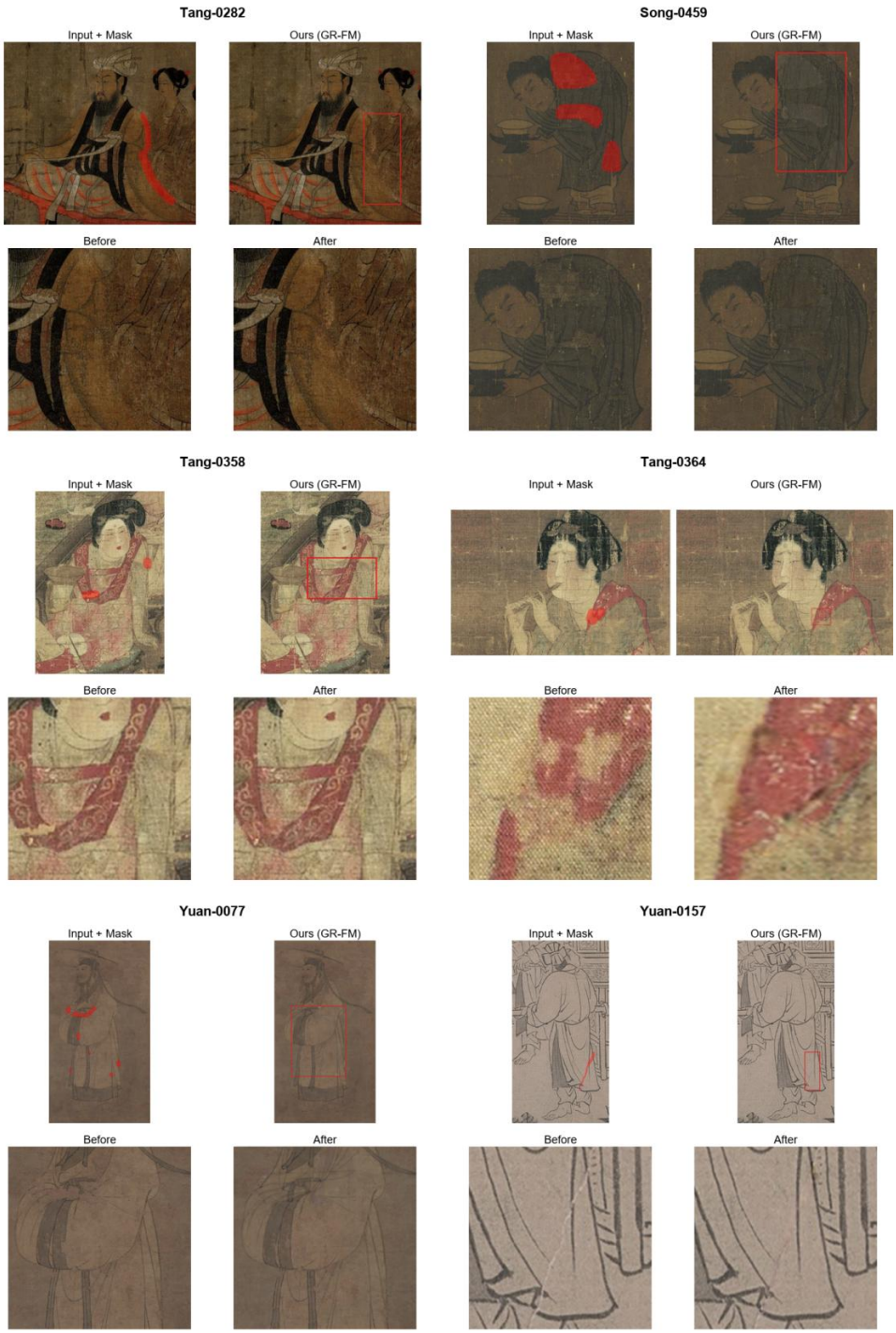


Fig. 8 | Qualitative validation of the proposed method on genuinely damaged garment regions in ancient paintings.

Figure 8 presents six representative cases. For narrow contour breaks such as Tang-0282 and Yuan-0157, the model continued the principal line direction using the surviving structure on both sides of the damage. In low-contrast cases such as Song-0459 and Yuan-0077, the restored regions remained relatively consistent with the surrounding color tones. For the larger irregular damaged regions in Tang-0358 and Tang-0364, the model reconstructed the overall color fields and portions of the garment edges, although multiple interpretations of local details remained plausible.

These cases show that the proposed method can generate visual restoration candidates within manually specified regions. However, this experiment does not establish the historical authenticity of the generated content. Practical applications must incorporate evidence from dress history, comparable images, and the judgment of conservation specialists.

Discussion

The proposed framework does not merely concatenate two modules; instead, it addresses two distinct limitations of the base inpainting model through stage-specific interventions. Fashion LoRA uses modern garment line drawings to initialize contours and folds, whereas GR-FM increases the loss weight of latent tokens in damaged regions, reducing the dilution of supervision for small, elongated defects by global errors.

The four controlled configurations show that the structural contribution of Fashion LoRA is expressed primarily through edge metrics in the modern-garment source domain, while its effect in the ancient-painting target domain is strongly metric dependent. Stylized drapery conventions, low-contrast brushwork, and differences in historical garment forms may cause source-domain structural biases to appear as either spurious or omitted edges. Therefore, Fashion LoRA should be interpreted as a structural initialization mechanism rather than as a consistently validated cross-domain adaptation method. GR-FM demonstrates more stable task adaptability across initialization conditions and significantly improves the pixel-level, structural, and perceptual metrics of Fashion-LoRA-initialized models. The sensitivity analysis further shows that BF1-2 approaches saturation when $\lambda_m \geq 10$, whereas pixel-level and perceptual metrics continue to improve gradually as the weight increases. The choice of $\lambda_m = 20$ represents a compromise between clear gains and a moderate reweighting strength. Hard-mask compositing directly preserves the original input outside the mask, providing restoration results with explicit, recordable, and reviewable spatial boundaries.

First-stage training on line drawings suppresses appearance-related factors in modern garment images, including figures, backgrounds, photographic lighting, and material highlights, allowing LoRA to focus more directly on garment line structure. At the same time, hard-mask compositing strictly confines generated content to damaged regions and preserves the original input outside the mask, improving spatial controllability and result traceability. These two design choices reduce uncertainty in generative restoration at the levels of structural initialization and spatial constraint, respectively, making the model more suitable for digitally assisted restoration.

The proposed method nevertheless has several limitations. Automatic edge extraction may introduce discontinuities, spurious edges, and residual textures, while modern garment line drawings cannot fully represent historical features such as wide robes with voluminous sleeves, complex layering, and stylized drapery conventions. In addition, the current experiments primarily simulate linear and localized irregular damage and do not systematically cover complex deterioration patterns such as insect damage or large-area pigment loss. Future work should incorporate line drawings from ancient

paintings and historical dress-form references, while expanding the range of deterioration masks and developing a hierarchical evaluation framework to strengthen target-domain structural guidance.

This study presents a two-stage framework for inpainting damaged garments in ancient paintings through Fashion LoRA structural guidance and GR-FM region-focused supervision. Fashion LoRA initializes contours and folds for color reconstruction, whereas GR-FM alleviates insufficient supervision for small and elongated damaged tokens. Four controlled configurations trained with three random seeds show that Fashion LoRA primarily improves edge metrics in the modern-garment source domain, while GR-FM provides more stable task adaptability across initialization conditions and exhibits significant interactions with Fashion LoRA for selected metrics. The complete two-stage configuration outperforms general-purpose baselines on most metrics in the ancient-painting evaluation set. The sensitivity analysis shows that BF1-2 approaches saturation for $\lambda_m \geq 10$, while pixel-level and perceptual metrics continue to improve gradually with stronger reweighting; $\lambda_m = 20$ is therefore adopted as a moderate setting. Hard-mask compositing further confines the modification region. The proposed method thus provides not only a practical approach to ancient-painting garment inpainting but also a controllable, verifiable, and extensible technical framework for generative restoration in cultural heritage.

Funding Declaration
This work was supported by the Shaanxi Provincial Social Science Fund Project, Project No. 2023G001.

References

1. Cornelis, B. *et al.* Crack detection and inpainting for virtual restoration of paintings: The case of the Ghent Altarpiece. *Signal Processing* **93**, 605–619 (2013).

2. Hou, M. *et al.* Virtual restoration of stains on ancient paintings with maximum noise fraction transformation based on the hyperspectral imaging. *Journal of Cultural Heritage* **34**, 136–144 (2018).

3. Ding, Q.-K. & Liang, H.-E. Digital restoration and reconstruction of heritage clothing: a review. *Herit Sci* **12**, 225 (2024).

4. Pietroni, E. & Ferdani, D. Virtual Restoration and Virtual Reconstruction in Cultural Heritage: Terminology, Methodologies, Visual Representation Techniques and Cognitive Models. *Information* **12**, 167 (2021).

5. Telea, A. An Image Inpainting Technique Based on the Fast Marching Method. *Journal of Graphics Tools* **9**, 23–34 (2004).

6. Bertalmio, M., Sapiro, G., Caselles, V. & Ballester, C. Image inpainting. in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques* 417–424 (ACM Press/Addison-Wesley Publishing Co., USA, 2000). doi:10.1145/344779.344972.

7. Criminisi, A., Perez, P. & Toyama, K. Region Filling and Object Removal by Exemplar-Based Image Inpainting. *IEEE Trans. on Image Process.* **13**, 1200–1212 (2004).

8. Morotti, E., Merizzi, F., Evangelista, D. & Cascarano, P. Inpainting with style: forcing style coherence to image inpainting with deep image prior. *Front. Comput. Sci.* **6**, 1478233 (2024).

9. Zhou, Z. *et al.* Inpainting Digital Dunhuang Murals with Structure-Guided Deep Network. *J. Comput. Cult. Herit.* **15**, 1–25 (2022).

10. Chen, Y., Chen, J., Ai, Y. & Tao, M. Mural image inpainting based on edge missing reconstruction and improved priority. *Laser Optoelectron. Prog.* **58**, 2410009 (2021).

11. Li, L. *et al.* Line Drawing-Guided Progressive Inpainting for Mural Damage. *J. Comput. Cult. Herit.* **18**, 48:1-48:20 (2025).

12. Sumathi, G. & Uma Devi, M. Inpainting of damaged temple murals using edge- and line-guided diffusion patch GAN. *Front. Artif. Intell.* **7**, 1453847 (2024) .

13. Lyu, Q., Zhao, N., Song, J., Yang, Y. & Gong, Y. Mural inpainting via two-stage generative adversarial network. *npj Herit. Sci.* **13**, 188 (2025).

14. Suvorov, R. *et al.* Resolution-robust Large Mask Inpainting with Fourier Convolutions. in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* 3172–3182 (2022). doi:10.1109/WACV51458.2022.00323.

15. Quan, W., Chen, J., Liu, Y., Yan, D.-M. & Wonka, P. Deep Learning-Based Image and Video Inpainting: A Survey. *Int J Comput Vis* **132**, 2367–2400 (2024).

16. Yu, J. *et al.* Generative image inpainting with contextual attention. In P*roc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 5505–5514 (2018).

17. Lugmayr A. *et al.* RePaint: Inpainting using denoising diffusion probabilistic models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 11461–11471 (2022).

18. Rombach R., Blattmann A., Lorenz D., Esser P. & Ommer B. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 10684–10695 (2022).

19. Saharia, C. *et al.* Palette: Image-to-Image Diffusion Models. in *ACM SIGGRAPH 2022 Conference Proceedings* 1–10 (Association for Computing Machinery, New York, NY, USA, 2022). doi:10.1145/3528233.3530757.

20. Zhang L., Rao A. & Agrawala M. Adding conditional control to text-to-image diffusion models. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 3836–3847 (2023).

21. Xie S., Zhang Z., Lin Z., Hinz T. & Zhang K. SmartBrush: Text and shape guided object inpainting with diffusion model. In *Proc. IEEE/CVF Conf. Comput. Vis.*

Pattern Recognit., 22428–22437 (2023).

22. Van Vijle, A., Hacıgüzeller, P. & Van der Snickt, G. Machine learning for painting conservation: a state-of-the-art review. *npj Herit. Sci.* **13**, 437 (2025).

23. Deng, X. & Yu, Y. Ancient mural inpainting via structure information guided two-branch model. *Herit Sci* **11**, 131 (2023).

24. Ma, Y. *et al.* OraGAN: a deep learning based model for restoring Oracle Bone Script Images. *npj Herit. Sci.* **13**, 681 (2025).

25. Chang, I.-C., Wun, Z.-S. & Yeh, H.-Y. An image inpainting technique on Chinese paintings. *J. Comput.* **29**, 121–135 (2018).

26. Nazeri, K., Ng, E., Joseph, T., Qureshi, F. & Ebrahimi, M. EdgeConnect: Structure Guided Image Inpainting using Edge Prediction. in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* 3265–3274 (2019). doi:10.1109/ICCVW.2019.00408.

27. Dong Q., Cao C. & Fu Y. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 11358–11368 (2022).

28. Li W. *et al.* MAT: Mask-aware transformer for large hole image inpainting. In Proc. *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 10758–10768 (2022).

29. Zeng, Y., Fu, J., Chao, H. & Guo, B. Aggregated Contextual Transformations for High-Resolution Image Inpainting. *IEEE Transactions on Visualization and Computer Graphics* **29**, 3266–3280 (2023).

30. Ren, N. & Huang, K. MAPGR: Multi-Agent Prompt-Guided Residual Diffusion for ancient mural restoration. *npj Herit. Sci.* **14**, 319 (2026).

31. Zhang, J., Bai, S., Zeng, X., Liu, K. & Yuan, H. Supporting historic mural image inpainting by using coordinate attention aggregated transformations with U-Net-based discriminator. *npj Herit. Sci.* **13**, 305 (2025).

32. Tang, X. *et al.* Digital restoration of mural paintings from late Tang tomb M1373 in Xi'an based on hyperspectral analysis and image interaction processing. *npj Herit. Sci.* **13**, 192 (2025).

33. Hu, J., Yu, Y. & Zhou, Q. GuidePaint: lossless image-guided diffusion model for ancient mural image restoration. *npj Herit. Sci.* **13**, 118 (2025).

34. Xue, J., Wang, X., Zhang, Q., Yang, X. & Li, C. Generating archaeological line drawings from limited reference images. *npj Herit. Sci.* **14**, 247 (2026).

35. Tian, R., Wu, T., Feng, D., Xin, Z. & Wang, L. Edge-guided interactive fusion of texture and geometric features for Dunhuang mural image inpainting. *Displays* **92**, 103297 (2026).

36. Jia, M. *et al.* Automatic Restoration of Dunhuang Murals and Process Visualization Method Based on Deep Learning. *Appl. Sci.* **15**, 1422 (2025).

37. Ge, Y., Zhang, R., Wang, X., Tang, X. & Luo, P. DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 5337–5345 (2019).

38. Saranya, M. S. & Geetha, P. Cross-domain fashion cloth retrieval via novel attention-guided cascade neural network and clothing parsing. *Computer Vision and Image Understanding* **235**, 103777 (2023).

39. Ren, H., Zhang, C., Li, Z. & Su, Z. EdgeConnect model based on Fourier convolution for Dunhuang mural costume image restoration. *npj Herit. Sci.* **13**, 592 (2025).

40. Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M. & Le, M. Flow matching for generative modeling. In *International Conference on Learning Representations* (2023).

41. Liu, X., Gong, C. & Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations* (2023).

42. Xiao B. *et al.* Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 4818–4829 (2024).

43. 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).

44. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 586–595 (2018).

45. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* **6**, 65–70 (1979).

46. Podell, D. *et al.* SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. in International Conference on Learning Representations (2024).

47. Ju, X. *et al.* BrushNet: A Plug-and-Play Image Inpainting Model with Decomposed Dual-Branch Diffusion. in *Computer Vision – ECCV 2024* (eds Leonardis, A. et al.) 150–168 (Springer Nature Switzerland, Cham, 2025). doi:10.1007/978-3-031-72661-3_9.