

Digital Reconstruction of Lacquer Art Surfaces Using Computer Vision and Generative Models

xuelan He

He.xuelan@outlook.com

Jiangxi Vocational and Technical College of Communications

Article

Keywords: lacquer art restoration, cultural heritage digitization, diffusion models, ControlNet, LoRA, structure-aware generation

Posted Date: June 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9988270/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Digital Reconstruction of Lacquer Art Surfaces Using Computer Vision and Generative Models

Xuelan He^{1*}

1.School of Information Engineering, Jiangxi Vocational and Technical College of Communications, NanChang, China

*Corresponding author email:913488089@qq.com (Xuelan He)

Abstract: The digital preservation of lacquer art faces a persistent challenge: existing methods often sacrifice structural fidelity for visual style transfer or neglect the layered materiality, crack networks, and characteristic gloss of lacquerware surfaces. This paper proposes the Structure-Aware Collaborative Generative Framework (SACGF), a task-oriented pipeline that integrates Stable Diffusion with Low-Rank Adaptation (LoRA) for domain-specific style injection, a self-trained, multi-condition ControlNet for specialized structural guidance, and the Segment Anything Model (SAM) for region-aware semantic constraint to prevent cross-region texture leakage. Lacquer surface reconstruction is formalized as a multi-objective optimization problem balancing structural consistency, texture fidelity, and style coherence. A purpose-built dataset of 2,194 high-resolution lacquer art images covering 8 major categories and 32 subcategories is constructed, with paired damaged–intact samples generated via degradation simulation. To complement subjective assessment with quantitative reproducibility, a Digital Lacquer Fidelity Score (DLFS) is introduced, integrating texture conformity, color fidelity, and structural similarity into a single composite score. Experimental results show that SACGF consistently outperforms CycleGAN, SD+LoRA, and progressively constrained ControlNet variants across perceptual and structural metrics, achieving a DLFS of 0.87 on the test set. Ablation studies further demonstrate that edge conditions primarily strengthen structural preservation, depth maps improve layer consistency, and SAM semantic masks reduce cross-region color and texture leakage. The proposed framework offers a practical and reproducible approach for structure-aware lacquer-art restoration and demonstrates a quantifiable pathway from visual resemblance toward cultural fidelity.

Keywords: lacquer art restoration; cultural heritage digitization; diffusion models; ControlNet; LoRA; structure-aware generation

1. Introduction

1.1 Background and Motivation

Lacquer art, as a precious treasure of China's intangible cultural heritage, embodies thousands of years of aesthetic heritage and craftsmanship [1, 2]. Its unique artistic language—characterised by intricate layered structures, exquisite crackle patterns (duanwen), and a warm lustre that shifts with the play of light and shadow—represents not only the pinnacle of decorative art but also serves as a tangible vessel for cultural memory and artisanal knowledge [3]. From the vermilion-painted black lacquer of the Warring States period to the carved lacquer and lacquer-inlaid woodwork of the Ming and Qing dynasties, every crack on the surface of a lacquerware piece and every layer of lacquer applied bears witness to the passage of time and the dialogue between craftsman and material. However, this precious cultural heritage is currently facing a severe crisis of transmission.

The creation and restoration of lacquer art rely heavily on tacit knowledge passed down through oral and practical instruction between master and apprentice; the process of mastering these skills is both lengthy and extremely challenging. As the older generation of artisans gradually retires, the number of successors capable of undertaking high-standard lacquerware restoration and creation has plummeted [4]. At the same time, extant lacquerware artefacts are suffering irreversible damage—such as fading pigments, flaking lacquer layers and cracking of the base material—due to natural ageing and environmental erosion. This raises the question of how digital technology can be used to support high-fidelity preservation and virtual restoration of lacquerware surfaces, focusing on key material and textural characteristics, rather than solely visual documentation.

Digital technology is playing an increasingly crucial role in the field of cultural heritage conservation [5]. Against this backdrop, AI-based image generation and restoration technologies have opened up entirely new avenues for the digital preservation of lacquer art [6]. The primary task is to generate reconstructed images from damaged or incomplete inputs, aiming to preserve semantic content, geometric structure, and color distribution, while approximating the texture and stylistic characteristics of lacquer art [7]. This technology not only provides cultural heritage conservators with reversible virtual restoration simulations, enabling them to carefully evaluate the effects of different intervention strategies before undertaking physical restoration; it also supports immersive digital museum experiences, allowing the public to appreciate the minute details of lacquerware in a virtual space; furthermore, it offers contemporary lacquer artists a vast array of generative patterns and textures for reference, stimulating the innovative vitality of traditional craftsmanship within a contemporary context. Nevertheless, given the complex layering, optical variations, and intricate crack networks of lacquerware surfaces, achieving reconstruction outcomes that faithfully reflect the material and aesthetic characteristics of lacquer art—what we term cultural fidelity—remains a challenging task.

1.2 Problem Statement and Research Gaps

At present, methods for the digital simulation and reconstruction of lacquer surfaces can be broadly categorised into two main paradigms; however, both have revealed fundamental limitations when dealing with the complex textures of lacquer.

The first category comprises physics-based simulation methods. These methods seek to reproduce the visual appearance of a paint finish by modelling physicochemical processes such as paint film curing, multi-layer coating and drying shrinkage [8]. Although they can generate images with a certain degree of realism under specific conditions, these methods entail extremely high computational costs and rely heavily on manual parameter tuning based on experience. More crucially, the rigid structure of physical models struggles to capture the flexible, organic and intuitive dynamic responses demonstrated by master craftsmen when dealing with material flow and environmental changes; consequently, the results produced often fall into the trap of being ‘technically accurate’ yet ‘artistically rigid’ [9].

The second category comprises generative methods based on deep learning, primarily including feature decoupling methods based on convolutional neural networks (CNNs) (such as AdaIN) and domain transfer methods based on generative adversarial networks (GANs) (such as CycleGAN) [10, 11]. These methods have achieved remarkable results in style transfer for general natural images. However, when applied to artworks such as lacquer art, which possess highly complex, multi-scale textural features, the challenges become particularly pronounced. Firstly, methods based on global statistical matching (such as the Gram matrix or the mean-variance

alignment in AdaIN) tend to generate smooth textures [12], making them highly prone to losing crucial micro-details in the lacquer surface—such as the sharp edges of crackle patterns, the subtle sense of thickness in lacquer layer accumulation, and the lustre that varies with viewing angle. Secondly, although attention mechanisms have been introduced to enhance the fidelity of local details, these models remain inherently data-hungry [13]. Although this work has constructed a lacquer art dataset comprising over a thousand images, this remains a relatively small-to-medium-sized dataset compared to the millions of images in natural image datasets; consequently, parameter-efficient fine-tuning strategies such as LoRA have been proposed [14]. However, when such models are applied to highly specialised cultural heritage datasets—which are limited in scale, domain-specific, and characterised by long-tail features—their generalisation capabilities often decline significantly [15]. Thirdly, in the absence of paired data, GAN-based methods are prone to ‘distorting’ the contours or semantic structures of objects during domain transfer, resulting in unacceptable artefacts [16]. This constitutes a fatal flaw for cultural heritage restoration, which demands that ‘restoration should be as close to the original as possible’.

Through a critical examination of existing approaches, a fundamental research gap emerges: no single model or unified framework can simultaneously ensure high fidelity in three core dimensions—structural consistency (precise retention of cracks and contours), textural authenticity (complex textures and rich details of layered materials), and stylistic coherence (the aesthetic and cultural essence of lacquer art) [17]. Existing studies either prioritize visual plausibility while neglecting material texture, or focus on texture synthesis at the cost of structural control [18]. More critically, in AI-driven discussions on cultural heritage restoration, the core value of “cultural authenticity” has long lacked a quantifiable evaluation system and actionable technical pathway. Consequently, many dazzling results may look “authentic” visually, yet risk becoming “sophisticated forgeries” in terms of craftsmanship and historical accuracy [19].

To bridge the above critical gap, this paper proposes a Structure-Aware Collaborative Generative Framework (SACGF) for the digital surface reconstruction of lacquer art. The core idea is that high-fidelity cultural reconstruction cannot rely on a single “black-box” model; instead, a collaborative system is required to jointly inject structure, texture, and semantics as explicit priors—mutually constraining and supporting each other—into a controllable generative process. Built upon the powerful generative backbone of Stable Diffusion (SD), SACGF integrates three functional modules: a Low-Rank Adaptation (LoRA) for accurately injecting lacquer style on medium-sized datasets, our purpose-trained, multi-condition ControlNet for providing precise structural and color constraints, and a Segment Anything Model (SAM) for fine-grained semantic guidance.

The logic of this architecture rests on a core hypothesis: only by taking the physical structure (crack networks, contour edges) and semantic layout (lacquer surface regions, decorative patterns) of lacquer art as explicit conditions to force a diffusion model already fine-tuned on lacquer-specific texture distributions, can I generate results that are both faithful to the input original and deeply imbued with the aesthetic qualities of lacquer art.

The main contributions of this paper are summarized as follows:

- (1) A structure-aware collaborative framework for digital restoration of lacquer surfaces, SACGF. Built upon Stable Diffusion, this framework integrates LoRA for style adaptation, ControlNet for structural constraints, and SAM for semantic priors, working collaboratively to achieve high-fidelity restoration of damaged lacquer images, balancing structural consistency, textural

authenticity, and stylistic coherence.

(2) A data construction and training mechanism tailored for lacquer restoration tasks.

To address the lack of restoration-oriented samples in existing lacquer datasets (which primarily serve classification, retrieval, and generative design), I construct a dataset of 2,194 images covering 8 major categories and 32 subcategories. Paired damaged-intact samples are generated via degradation simulation, providing a training foundation for diffusion models to learn crack completion, lacquer layer repair, and surface reconstruction.

(3) A Digital Lacquer Fidelity Score (DLFS) system for cultural authenticity in digital lacquer art. I propose a comprehensive evaluation system centered on the Digital Lacquer Fidelity Score (DLFS). This system quantifies three key dimensions: texture conformity (combining LPIPS and GLCM statistics), color fidelity (CIEDE2000-based color difference and color space distribution consistency), and structural similarity (integrating SSIM, multi-scale SSIM, and edge retention). It provides an objective, reproducible standard for evaluating AI-generated lacquer images, going beyond subjective perception.

1.3 Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews related work on digital restoration of cultural heritage, generative models, and material characterization of lacquer art, further clarifying the research positioning of this study. Section 3 presents the mathematical principles and technical details of the proposed SACGF framework, including the formal problem definition, structure-aware feature extraction, generative reconstruction mechanism, and training strategy. Section 4 describes the experimental setup, dataset construction, and evaluation metrics, followed by comprehensive assessment results from quantitative comparisons, visual comparisons, and ablation studies. Section 5 offers an in-depth discussion of the implications of our findings, focusing on the deeper meaning of cultural authenticity, ethical considerations, and the limitations of the current approach. Finally, Section 6 concludes the paper and outlines directions for future research.

2. Related Work

The digital reconstruction of lacquer surfaces is a highly interdisciplinary topic, situated at the intersection of computer vision, generative models, materials science, and cultural heritage preservation. To clarify the positioning and contributions of this study, I review related work from three progressive dimensions. First, I trace the craftsmanship traditions and material properties of lacquer art, which form the cognitive foundation for any digital effort. Second, I survey digital preservation and restoration techniques for cultural heritage, analyzing the strengths and limitations of general-purpose methods. Finally, I focus on AI-driven generation and design innovation in lacquer art, discussing the potential and bottlenecks of cutting-edge generative models in this specific domain [20]. This review highlights a core gap: existing methods do not provide an end-to-end framework that integrates material structure, micro-texture, and stylistic characteristics for lacquer art restoration.

2.1 Craftsmanship Traditions and Material Properties of Lacquer Art

Lacquer art, historically known as *xiushi* (decorative lacquerwork), traces its origins to the Neolithic period. A red-lacquered wooden bowl unearthed at the Hemudu site in Yuyao, Zhejiang, provides evidence that ancient Chinese artisans had mastered the use of natural lacquer as early as

7,000 years ago [21,22]. Over thousands of years, the materials, techniques, and artistic styles of lacquerware have continuously evolved. By the Ming dynasty, Huang Cheng's craft treatise *Xiushilu* (The Treatise on Lacquer) had systematically documented over 400 decorative techniques, marking the full maturity of ancient China's lacquer craftsmanship system [23,24,25].

To understand lacquer art, one must start from its unique materiality. Natural raw lacquer is a milky white sap tapped from lacquer trees, composed primarily of urushiol, laccase, gum, and water. It solidifies into a film under specific temperature and humidity conditions, earning its reputation as the “king of coatings”. The core material characteristics, which pose the greatest challenges for digital simulation, are as follows (see Figure 1):

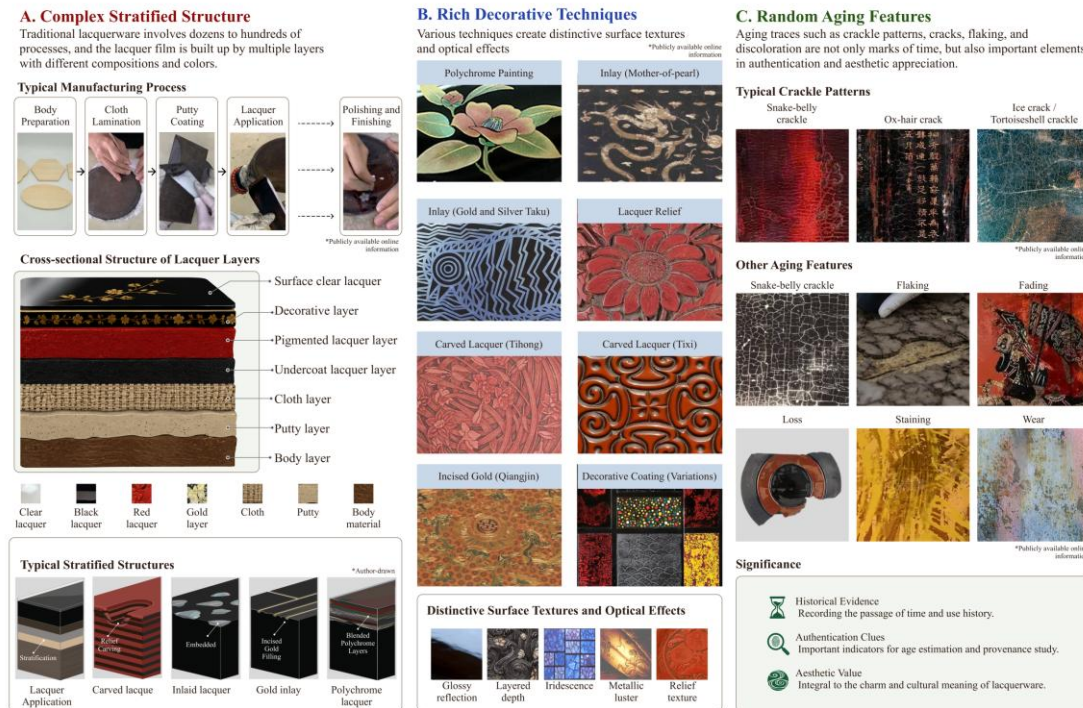


Fig1. Material Characteristics of Traditional Chinese Lacquer Art(Adapted and redrawn by the author based on publicly available workflow references.)

(i) Complex layered structures: Traditional lacquerware fabrication involves dozens or even hundreds of processes, including body forming, fabric backing, ash scraping, and lacquer brushing. The resulting film is built up from multiple layers of coatings with different compositions and colors [26].

(ii) Rich decorative techniques: Methods such as painting, inlay (mother-of-pearl, gold/silver flat inlay), piled lacquer, carved lacquer (tihong, tixi), incised gold, and bentuyou (free-style marbling) each create unique surface textures and optical effects [27,28].

(iii) Stochastic aging features: Excavated or historic lacquerware surfaces often exhibit crackle patterns (e.g., snake-belly cracks, ox-hair cracks), fissures, flaking, and fading. These traces of age are not merely signs of deterioration but integral components of authentication and aesthetic appreciation [25,29].

Therefore, the digitalization of lacquer art extends far beyond what a photograph or a 3D model can capture. It demands an accurate interpretation and reproduction of material, craftsmanship, and history.

2.2 Digital Preservation and Restoration of Cultural Heritage

Digital preservation of cultural heritage has developed into a technical system based on 3D scanning, photogrammetry, and image processing [30]. These technologies can efficiently record the geometric morphology and surface textures of artifacts without physical contact, and are widely used for documentation of ancient architecture, cave temples, and movable cultural relics.

In the field of lacquer art and painting, digital technologies have evolved along two main directions:

Auxiliary analysis: Techniques such as X-ray, multispectral photography, and reflectance transformation imaging (RTI) are used to reveal the internal structure, underlying drafts, and deterioration conditions of lacquerware, providing a scientific basis for restoration planning [31].

Image enhancement and virtual restoration: Researchers have begun to introduce intelligent algorithms to restore the visual quality of low-light or damaged lacquer paintings. For example, Wang et al. proposed a method combining the sparrow search algorithm (SSA) with an incomplete beta function for adaptive enhancement of lacquer painting images under low-light conditions, demonstrating the potential of optimization algorithms in image enhancement tasks [32]. However, such methods mainly operate at the level of global pixel-wise adjustments. They lack a semantic understanding of image content (e.g., cracks, patterns) and cannot achieve structural-level reconstruction or repair.

A limitation is that most methods operate on 2D images, while lacquerware's visual and tactile qualities depend on its 3D layered structure. They cannot simulate the complex refraction and reflection of light across lacquer layers, inlays, or carved edges, let alone reproduce its tactile warmth.

2.3 AI-Driven Generation and Design Innovation for Lacquer Art

In recent years, AI generative models, particularly deep learning, have brought paradigm-level innovations to the digital preservation of cultural heritage [33]. Starting from general methods for image restoration and style transfer, this subsection focuses on their application to the specific domain of lacquer art, comparing and analyzing the strengths and limitations of different technical approaches.

2.3.1 Deep Learning-Based Image Restoration and Style Transfer

Image restoration and style transfer are fundamental tools for digital reconstruction. Early CNN-based methods exploit hierarchical features to separate content from style but lack the ability to control structure [34]. Subsequently, generative adversarial networks (GANs), represented by CycleGAN, achieve unpaired style transfer. However, their training is unstable, and they are prone to generating “artifacts” that do not conform to the input structure, making them unacceptable in the cultural heritage domain, where authenticity is paramount [35]. In recent years, diffusion models, represented by Stable Diffusion, generate high-quality images through iterative denoising [36], surpassing GANs in both detail richness and content consistency. More importantly, modules such as ControlNet allow spatial structure information (e.g., line drawings, depth maps) to be injected as conditioning into the generation process, enabling precise control over geometric contours [37]. Meanwhile, parameter-efficient fine-tuning techniques such as LoRA make it possible to customize specific styles on limited datasets. These advances provide the key technical foundation for high-fidelity digital reconstruction of lacquer surfaces.

2.3.2 Applications of Generative Models in Cultural Heritage and the Lacquer Art Case

Generative models have been applied to tasks such as Dunhuang mural restoration, embroidery, and ceramic pattern generation [38,39,40]. These works provide important methodological references for this study.

Specifically in the lacquer art domain, application explorations are still in their early stages but have shown significant potential. In the book *Chushi Qiqi* (Chu-Style Lacquerware) compiled by Zhao Yuchun, a large collection of plates systematically presents the mysterious shapes and brilliant decorative patterns of Warring States-period Chu-style lacquerware. Such ancient texts themselves serve as important stylistic corpora [26]. In terms of digital practice: Bi et al. constructed the first dedicated Chu-style lacquerware image dataset and fine-tuned a generative model using LoRA, demonstrating the decisive role of domain-specific datasets in improving the generation quality of particular lacquer styles [41]. Wang et al. applied improved Canny and HOG features to extract and recognize lacquerware patterns, and then optimized geometric forms using a BP-GA (genetic algorithm). Their core focus was on the contour and formal aesthetics of the patterns [42]. Liang et al. proposed a “human-in-the-loop” lacquerware design system (HITL-AIGC), integrating a diffusion model, LoRA, and ControlNet, allowing artisans to guide AI generation through text and sketches, marking an important step in human-AI collaborative innovation [43].

The above work primarily treats lacquer art as visual style or decorative pattern, which differs from the layered materiality and structural features discussed in Section 2.1: the beauty of lacquerware lies not only in the image it presents, but also in how it was layered, carved, and cracked. This gap leads directly to the research question of this paper: can a generative framework be constructed that, starting from an image of a lacquer surface, reconstructs not only its style but also its material texture and structure?

2.4 Research Gap and Positioning of This Paper

Synthesizing the above review, the contributions of existing work are clear: from two-dimensional image enhancement [32] to deep learning-based style transfer, and from the CSL dataset [41] to the HITL-AIGC system [43], these studies demonstrate the potential of the technical pathway combining diffusion models, fine-tuning, and conditional control. However, their task objectives are all aimed at creating or designing new, complete lacquer artworks, with “patterns” and “colors” as the primary learning targets. The essential values of lacquer art—the thickness of multi-layer lacquer-ash buildup, the alternating textures at the cuts of carved lacquer, the sharp depth of natural crackle patterns, and the layered luster of gold powder beneath translucent lacquer—are systematically overlooked in these creation-oriented frameworks. Figure 2 illustrates several representative failure cases observed when conventional restoration algorithms are directly applied to lacquer surfaces.

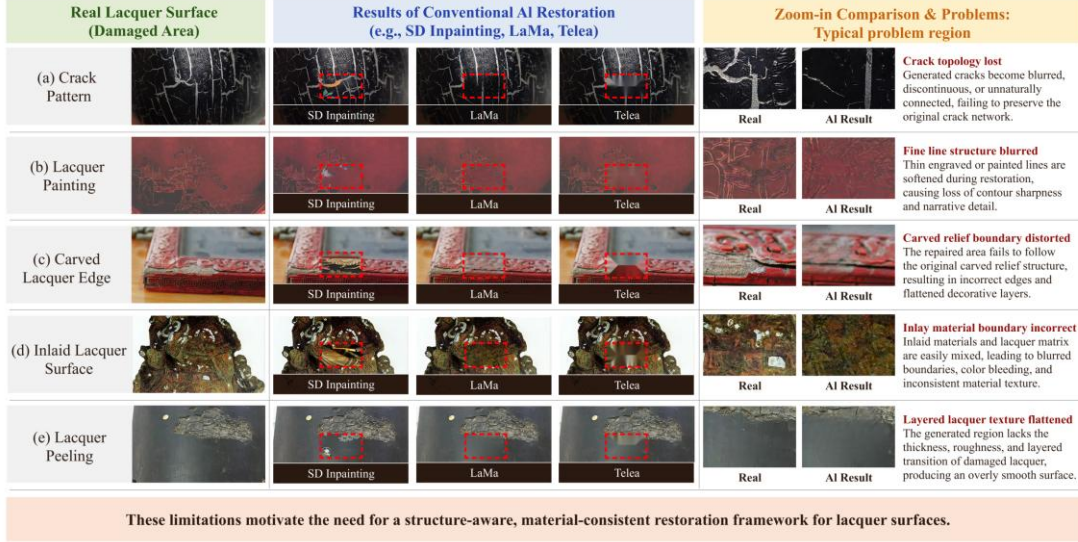


Fig2. Challenges in Digital Lacquer Surface Restoration

2.4.1 Comparison with Existing Digital Restoration Methods

Prior work in digital restoration has explored a wide range of techniques, from traditional inpainting algorithms that gradually fill missing regions using variational and patch-based schemes to deep learning-based approaches that learn content priors from data. For example, multi-layer inpainting techniques aim to restore textured artworks by considering hierarchical visual features, but lack explicit structural or semantic guidance [44]. Recent advances in semantic- and sketch-guided diffusion models for ancient paintings demonstrate fine-grained control over brushstroke details and artistic style, yet these methods primarily address planar painted surfaces and do not explicitly incorporate layered material structures or crack network priors, which are essential for lacquer art restoration [45]. Similarly, diffusion-based mural restoration techniques use line drawing guidance to preserve global structure, but such methods are tailored for large continuous surface repair and are less effective at capturing the rich micro-texture and multiple overlay layers found in lacquer surfaces [46]. In contrast, the proposed SACGF framework integrates multi-condition structural guidance (edge, depth, semantic mask) with style adaptation to jointly address structure consistency, texture fidelity, and style coherence—properties that are overlooked or underemphasized in existing restoration models.

As summarized in Table 1, existing methods either lack explicit structural guidance, semantic control mechanisms, or the ability to model layered lacquer-specific surface characteristics. These limitations motivate the development of a dedicated restoration framework for lacquer art.

Table 1. Comparison of Representative Digital Restoration Methods and SACGF

Method	Task	Structural Guidance	Semantic Control	Dataset Type	Limitations
Classical Inpainting	General inpainting	×	×	Natural images	No semantics
SSGR (Ancient Painting)	Painting	✓	✓ (semantic)	Painting	No layered structure
DiffInpaint (Murals)	Mural repair	✓ (line drawing)	×	Mural	Not lacquer surface
Proposed SACGF	Lacquer art	✓ (multi-condition)	✓ (SAM)	Lacquer dataset	—

Consequently, no framework to date has been dedicated to the digitization of lacquer surfaces specifically for restoration and reconstruction. This paper is intended to fill that gap. The goal of this

study is not to generate an image that merely possesses a lacquer-art style, but rather, starting from an image of a (possibly damaged) lacquer surface, to reconstruct its layered structure, material texture, and craft features with high fidelity. To this end, I propose the Structure-Aware Collaborative Generative Framework (SACGF), which integrates Stable Diffusion, LoRA, ControlNet, and SAM into a unified system, and design a Digital Lacquer Fidelity Score (DLFS) evaluation system that addresses the three dimensions of texture, color, and structure. This paradigm shift from “style creation” to “surface reconstruction”—that is, from “generating an image that looks like a lacquer painting” to “reconstructing a surface that feels like a lacquer object”—constitutes the fundamental innovation that distinguishes this paper from existing research.

3. Methodology

This section systematically presents the mathematical principles and technical details of the proposed Structure-Aware Conditional Generative Framework (SACGF) for digitizing lacquerware surfaces. It first provides a formal problem definition, then an overview of the framework, followed by in-depth discussions on structure-aware feature extraction, diffusion-based generative reconstruction, training strategies, evaluation protocols, dataset construction, and experimental setup.

3.1 Problem Formalization

The task is formulated as conditional image generation. Given a damaged or to-be-stylized input image $I_{in} \in \mathbb{R}^{H \times W \times 3}$, the objective is to generate a reconstruction that $I_{out} \in \mathbb{R}^{H \times W \times 3}$ approximates the input’s structural, textural, and color properties.

Let $\mathcal{X}$ be the space of damaged lacquerware images, $\mathcal{Y}$ the space of intact lacquerware images, and $\mathcal{F}$ the space of lacquerware style features. Find a mapping $\Phi: \mathcal{X} \rightarrow \mathcal{Y}$ such that for any input $I_{in} \in \mathcal{X}$, the output $I_{out} = \Phi(I_{in})$ optimizes the following three objectives:

1) Structure Consistency

Quantifies the alignment of geometric structure and semantic layout between output and input. The structural loss is:

$$\mathcal{L}_{struct}(I_{in}, I_{out}) = 1 - SSIM(I_{in}, I_{out}) + \lambda_e \cdot |\mathcal{E}(I_{in}) - \mathcal{E}(I_{out})|_2^2 \quad (1)$$

Where $SSIM(\cdot, \cdot)$ denotes the structural similarity index, $\mathcal{E}(\cdot)$ the Canny edge detector, and λ_e the edge fidelity weight. This term enforces strict correspondence of contours, crack patterns, and layer stratification, preventing structural semantic distortion.

2) Texture Authenticity

Measures the richness and realism of lacquerware texture in the output. The texture authenticity metric is:

$$\mathcal{L}_{tex}(I_{out}) = \lambda_l \cdot LPIPS(I_{out}, I_{gt}) + \lambda_{glcm} \cdot d_{GLCM}(T(I_{out}), T_{lacquer}) \quad (2)$$

Where $LPIPS(\cdot, \cdot)$ is the learned perceptual image patch similarity, $d_{GLCM}(\cdot, \cdot)$ the Frobenius distance in gray-level co-occurrence matrix (GLCM) feature space, $T(\cdot)$ a texture feature extractor, $T_{lacquer}$ the reference lacquer texture statistics, and $\lambda_l, \lambda_{glcm}$ balancing coefficients. This objective ensures the output exhibits the characteristic cracks, layering, and gloss of lacquerware.

3) Style Consistency

Measures the alignment between the generated image and the visual style statistics of lacquer art, including color distribution, texture rhythm, and decorative appearance. The style consistency

loss is defined as:

$$\mathcal{L}_{style}(\mathbf{I}_{out}, \mathcal{F}) = \sum_{l=1}^L \frac{1}{4N_l^2 M_l^2} \sum_{i,j} \left(G_{ij}^l(\mathbf{I}_{out}) - G_{ij}^l(\mathcal{F}) \right)^2 \quad (3)$$

Where $G_{ij}^l(\cdot) = \sum_k F_{ik}^l F_{jk}^l$ is the Gram matrix of the l -th feature map, F_{ik}^l the activation of the i -th filter at position k in layer l , N_l the number of filters, and M_l the product of spatial dimensions. By matching multi-scale feature statistics, this term encourages the generated image to approximate the visual appearance of traditional lacquer surfaces in color, texture, and stylistic coherence.

The overall optimization integrates the above losses into a unified multi-objective problem:

$$\min_{\Phi} \quad \mathcal{J}(\Phi) = \mathcal{L}_{struct} + \mathcal{L}_{tex} + \mathcal{L}_{style} + \mathcal{R}(\Phi) \quad (4)$$

Where $\mathcal{R}(\Phi)$ is a regularization term that prevents the model from overfitting to specific degradation patterns in the training data.

3.2 Framework Overview

To solve the multi-objective optimization problem defined in Equation (4), this section proposes the Structure-Aware Collaborative Generative Framework (SACGF). The framework combines Stable Diffusion (SD) for generation, Low-Rank Adaptation (LoRA) for style fine-tuning, ControlNet for structural conditioning, and SAM for semantic guidance, forming a pipeline from feature extraction and conditional reconstruction to multi-dimensional evaluation. Figure 3 illustrates the overall architecture of SACGF. The framework is organized along the following six-stage pipeline:

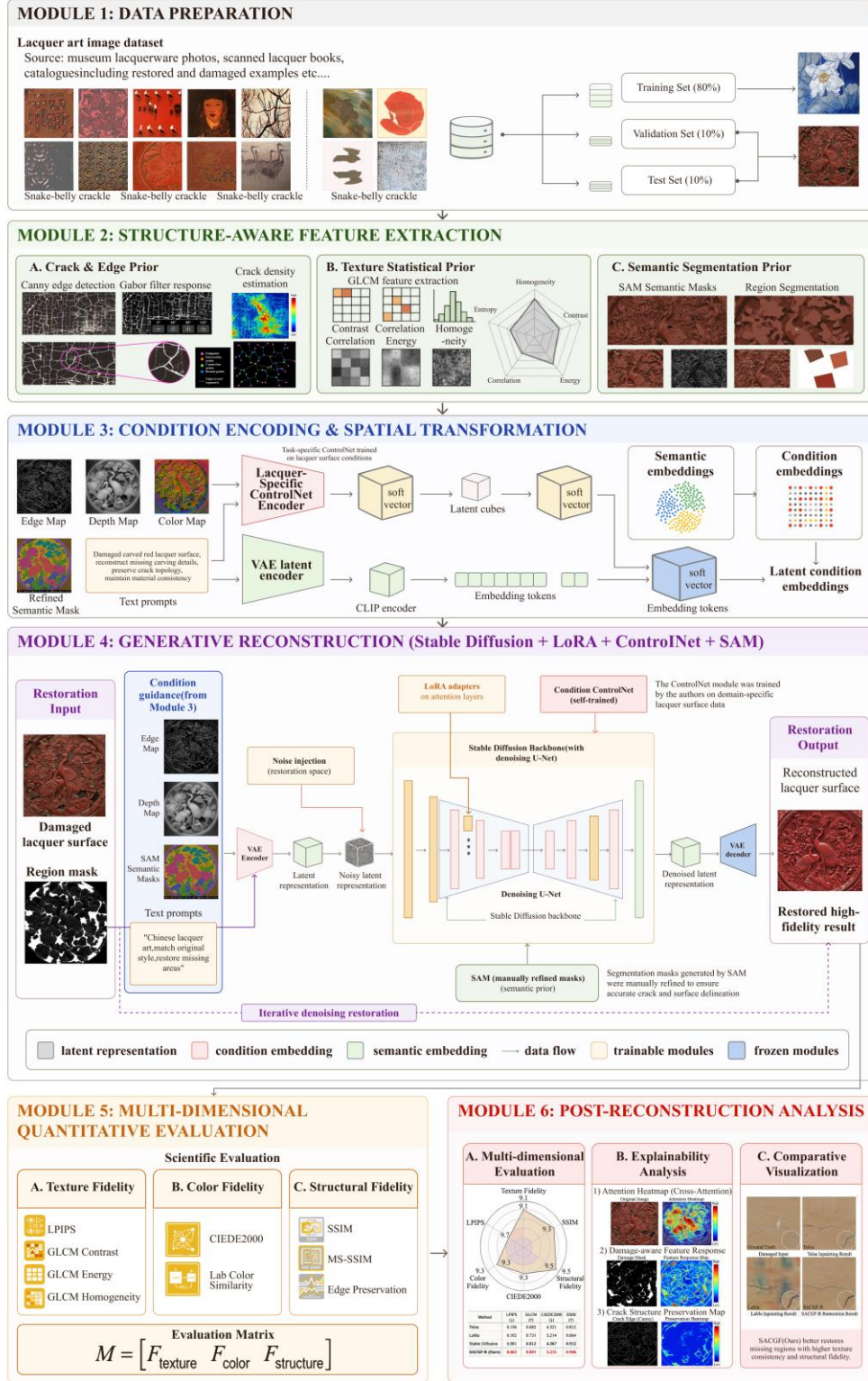


Fig3. Overall architecture of the proposed SACGF pipeline for lacquer surface reconstruction

1) Structure-Aware Feature Extraction

The input damaged lacquerware image first passes through a structure-aware feature extraction module, which extracts three types of complementary prior information in parallel:

1. Crack and edge prior: the crack network and geometric contours of lacquerware are extracted using multi-scale Canny edge detection and crack-density estimation.

2. Texture statistical prior: multi-scale texture descriptors are extracted using a pre-trained CNN (VGG-19) and gray-level co-occurrence matrices (GLCM).

3. Semantic segmentation prior: fine-grained semantic segmentation is performed by the SAM model, producing semantic region masks.

2) Condition Encoding & Spatial Transformatio

The structural priors and semantic masks extracted in Stage 1 are encoded into condition tensors and mapped to the latent space of the diffusion model. Specifically: our self-trained ControlNet, equipped with zero-convolution layers trained specifically for lacquer surfaces, applies condition encoding to edge maps, depth maps, and color maps; SAM masks are transformed into spatial constraints in the latent space via the VAE encoder; semantic labels are converted into semantic condition embeddings through the CLIP text encoder.

3) Guided Restoration

Guided by the condition tensors, an iterative denoising process is executed based on the UNet architecture of Stable Diffusion v1.5. The LoRA adapter injects lacquerware style parameters in a low-rank form, the ControlNet module provides structural constraints, and the SAM masks delineate semantic region boundaries. Starting from random noise, this stage generates a high-fidelity lacquerware reconstruction image after T denoising steps.

4) Multi-Dimensional Quantitative Evaluation

The generated image undergoes systematic quantitative evaluation, computing an indicator matrix across three core dimensions:

$$\mathbf{M} = [m_{tex}, m_{color}, m_{struct}]^T \in \mathbb{R}^{3 \times k} \quad (5)$$

Where m_{tex} denotes texture conformity (LPIPS, GLCM contrast, GLCM correlation, GLCM energy, GLCM homogeneity), m_{color} m_{struct} denotes structural similarity (SSIM, multi-scale SSIM, edge fidelity).

3.3 Structure-Aware Feature Extraction

The structure-aware feature extraction module is the core innovation that distinguishes the SACGF framework from conventional image generation approaches. The uniqueness of lacquerware surfaces lies in their stratified structures, crack networks, and multi-layered gloss variations—features that cannot be effectively captured by generic CNN features. This section proposes a multi-scale, multi-modal feature extraction strategy that comprehensively characterizes the visual properties of lacquerware surfaces from three dimensions: edge structure, texture statistics, and semantic regions.

3.3.1 Crack and Edge Structure Extraction

Let the input image I_{in} first be converted to a grayscale image $I_{gray} = 0.299R + 0.587G + 0.114B$. Crack structures exhibit multi-scale characteristics; therefore, a multi-level Gaussian smoothing and adaptive thresholding strategy is adopted.

Multi-scale Canny edge detection performs Gaussian smoothing and gradient computation at $\mathcal{S} = \{\sigma_1, \sigma_2, \dots, \sigma_K\}$ scales:

$$G_x^{(k)} = \frac{\partial}{\partial x} \left(I_{gray} * \frac{1}{2\pi\sigma_k^2} \exp\left(-\frac{x^2 + y^2}{2\sigma_k^2}\right) \right) \quad (6)$$

$$G_y^{(k)} = \frac{\partial}{\partial y} \left(I_{gray} * \frac{1}{2\pi\sigma_k^2} \exp\left(-\frac{x^2 + y^2}{2\sigma_k^2}\right) \right) \quad (7)$$

The gradient magnitude and orientation maps are:

$$M^{(k)}(x, y) = \sqrt{G_x^{(k)}(x, y)^2 + G_y^{(k)}(x, y)^2}, \quad \Theta^{(k)}(x, y) = \text{atan2}(G_y^{(k)}, G_x^{(k)}) \quad (8)$$

For each scale, non-maximum suppression and double-threshold hysteresis tracking are applied to the gradient magnitude, yielding a binary edge map $E^{(k)}$. The multi-scale fused edge map is obtained by weighted summation:

$$E_{\text{fused}} = \sum_k w_k \cdot E^{(k)}, \quad w_k = \frac{\sigma_k}{\sum_{j=1}^K \sigma_j} \quad (9)$$

Crack density estimation is introduced to distinguish lacquerware cracks from generic edges. Based on the edge map, the local crack density is computed as:

$$\rho_{\text{crack}}(x, y) = \frac{1}{|\mathcal{N}_r(x, y)|} \sum_{(x', y') \in \mathcal{N}_r(x, y)} E_{\text{fused}}(x', y') \quad (10)$$

Where $\mathcal{N}_r(x, y)$ is the local neighborhood of radius r centered at (x, y) . Regions with high crack density are assigned higher structural weights in subsequent conditional control.

3.3.2 Crack Topology Analysis

Lacquerware crack patterns, such as snake-belly cracks, ox-hair cracks, and ice cracks, show distinct geometric characteristics. To describe these differences, skeleton extraction is used to obtain simplified crack structures from the edge map. These skeletonized structures are used for qualitative analysis and post-reconstruction comparison rather than as an independent training loss. This analysis helps examine whether the reconstructed crack patterns preserve directional continuity, density distribution, and local branching characteristics.

Let the crack skeleton graph of the input image be $\mathcal{G}_{\text{in}} = (\mathcal{V}_{\text{in}}, \mathcal{E}_{\text{in}})$ and that of the reconstructed image be $\mathcal{G}_{\text{out}} = (\mathcal{V}_{\text{out}}, \mathcal{E}_{\text{out}})$. The topology fidelity loss is defined as:

$$\mathcal{L}_{\text{topo}} = 1 - \frac{2 \cdot |\mathcal{M}(\mathcal{G}_{\text{in}}, \mathcal{G}_{\text{out}})|}{|\mathcal{E}_{\text{in}}| + |\mathcal{E}_{\text{out}}|} \quad (11)$$

where $\mathcal{M}$ denotes the maximum edge-matching function. By minimizing the topological discrepancy between the crack networks, this loss ensures that the topological characteristics of the crack patterns (parallel, interwoven, network-like) are accurately retained.

3.3.3 Multi-Angle Texture Feature Extraction

The mathematical characterization of lacquerware texture is comprehensively captured through orthogonal analysis from the following three perspectives.

A. Gray-Level Co-occurrence Matrix (GLCM) Statistics

The GLCM provides a second-order statistical description of texture. The image is quantized into $L_g = 64$ gray levels. For a given direction θ and distance d , the GLCM is computed as:

$$P_{\theta, d}(i, j) = \sum_{x, y} 1 [I_{\text{gray}}(x, y) = i \wedge I_{\text{gray}}(x + d \cos \theta, y + d \sin \theta) = j] \quad (14)$$

Where $1[\cdot]$ is the indicator function. The matrix is normalized:

$$P_{\text{norm}}(i, j) = \frac{P_{\theta, d}(i, j)}{\sum_{i, j} P_{\theta, d}(i, j)} \quad (15)$$

From P_{norm} , four classical statistical features are computed:

$$1. \text{ Contrast : } f_{\text{con}} = \sum_{i, j} |i - j|^2 \cdot P_{\text{norm}}(i, j)$$

$$2. \text{ Correlation : } f_{\text{cor}} = \sum_{i, j} \frac{(i - \mu_i)(j - \mu_j) \cdot P_{\text{norm}}(i, j)}{\sigma_i \sigma_j}$$

$$3. \text{ Energy : } f_{\text{ene}} = \sum_{i, j} P_{\text{norm}}(i, j)^2$$

$$4. \text{ Homogeneity : } f_{\text{hom}} = \sum_{i,j} \frac{P_{\text{norm}}(i,j)}{1+|i-j|}$$

These features are computed across four directions $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ and three distances $d \in \{1, 2, 3\}$ and then averaged, yielding a 12-dimensional GLCM texture feature vector $f_{\text{GLCM}} \in \mathbb{R}^{12}$.

B. Deep Perceptual Feature Extraction

Deep texture features are extracted using a VGG-19 network pre-trained on ImageNet to approximate perceptual characteristics of lacquer textures. The input image is fed into VGG-19, and the feature maps from layers conv3_2, conv4_2, conv5_2 are obtained:

$$F_{\text{VGG}}^l = \text{VGG}_l(I_{\text{in}}), \quad l \in \{\text{conv3_2, conv4_2, conv5_2}\} \quad (16)$$

For each layer's feature map, the channel-wise mean and variance are computed as texture perceptual statistics:

$$\mu_l^c = \frac{1}{H_l W_l} \sum_{x,y} F_{\text{VGG}}^l(x, y, c), \quad \sigma_l^c = \sqrt{\frac{1}{H_l W_l} \sum_{x,y} (F_{\text{VGG}}^l(x, y, c) - \mu_l^c)^2} \quad (17)$$

The statistics from the three layers are concatenated to form the deep texture feature vector $f_{\text{CNN}} \in \mathbb{R}^{C_3+C_4+C_5}$, where C_l is the number of channels in layer l .

The comprehensive texture feature vector is defined as:

$$f_{\text{tex}} = [f_{\text{Gabor}}; f_{\text{GLCM}}; f_{\text{CNN}}] \in \mathbb{R}^{D_{\text{tex}}} \quad (18)$$

3.3.4 Semantic Structure Segmentation

Lacquerware images contain multiple visually distinct regions (e.g., base layer, cracks, decorative patterns). SAM is used to generate masks for these regions, providing semantic guidance during conditional reconstruction.

SAM Semantic Segmentation: The SAM model comprises an image encoder E_{img} , a prompt encoder E_{prompt} , and a mask decoder D_{mask} . For an input image I_{in} , dense features are first extracted via the image encoder:

$$Z_{\text{img}} = E_{\text{img}}(I_{\text{in}}) \in \mathbb{R}^{H' \times W' \times C_z} \quad (19)$$

In the interactive segmentation mode, given N_p prompt points $\mathcal{P} = (x_i, y_i, t_i)_{i=1}^{N_p}$ ($t_i \in \{0, 1\}$ indicating foreground/background labels), the prompt encoder outputs:

$$Z_{\text{prompt}} = E_{\text{prompt}}(\mathcal{P}) \in \mathbb{R}^{N_p \times C_z} \quad (20)$$

The mask decoder combines image features and prompt features to produce a semantic mask probability map:

$$M_{\text{sem}} = D_{\text{mask}}(Z_{\text{img}}, Z_{\text{prompt}}) \in [0, 1]^{H \times W \times N_c} \quad (21)$$

Where N_c denotes the number of semantic categories (e.g., lacquer base layer, crack regions, metallic decoration, decorative patterns). A hard segmentation is obtained via the $\arg \max$ operation:

$$S(x, y) = \arg \max_{c \in \{1, \dots, N_c\}} M_{\text{sem}}(x, y, c) \quad (22)$$

Semantic-aware region partitioning converts the segmentation map S into N_c binary masks $B_{cc} = 1^{N_c}$:

$$B_c(x, y) = 1[S(x, y) = c], \quad c = 1, \dots, N_c \quad (23)$$

These region masks serve as spatial constraints in the subsequent generation process, helping preserve region boundaries and reduce undesired texture leakage during reconstruction.

While SAM provides automatic segmentation, preliminary experiments revealed that fully automatic masks often contain errors at crack boundaries and decorative inlays, which are critical for lacquer art. Therefore, all SAM-generated masks undergo a manual refinement stage guided by the author, who has been trained in lacquer craft recognition. This process corrects segmentation boundaries and merges/splits semantic categories (e.g., distinguishing 'gold-inlay' from 'crack region' where they spatially overlap). The refined masks serve as the final semantic prior $\mathbf{B}_c$ in Eq. (23) (see in figure 4). This human-in-the-loop step ensures that the semantic guidance aligns with both physical material boundaries and art-historical categories.

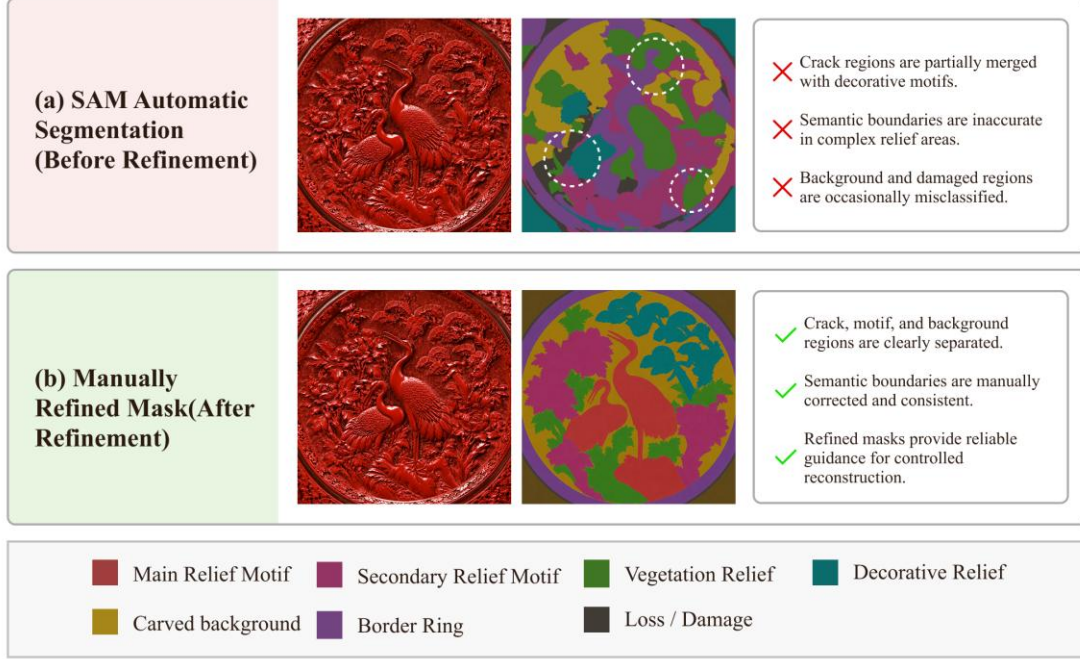


Fig4. Example of SAM mask refinement

3.4 Condition-Guided Inpainting and Reconstruction

The generative reconstruction module constitutes the core generative engine of the SACGF framework. Starting from the theoretical foundations of diffusion models, this subsection elaborates on the LoRA fine-tuning strategy, the multi-condition ControlNet guidance mechanism, and the integration of SAM semantic constraints.

3.4.1 Fundamentals of Stable Diffusion

Given a real lacquerware image sample $x_0 \sim q(x_0)$, the forward process gradually adds Gaussian noise over T time steps, forming a Markov chain:

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I\right) \quad (24)$$

Where $\beta_t \in (0,1)$ is a predefined noise schedule. By applying the reparameterization trick, x_t can be sampled directly from x_0 :

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \quad (25)$$

With $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$.

The reverse process learns to recover x_0 by iteratively denoising from pure noise $x_T \sim \mathcal{N}(0, I)$:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I) \quad (26)$$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(x_t, t) \right) \quad (27)$$

Where $\epsilon_\theta(x_t, t)$ is the UNet noise prediction network. Stable Diffusion shifts the diffusion process into a latent space: the pretrained VAE encoder $\mathcal{E}_{\text{VAE}}$ compresses the image into a latent representation $z_0 = \mathcal{E}_{\text{VAE}}(x_0) \in \mathbb{R}^{h \times w \times c}$ (typically $h = H/8, w = W/8, c = 4$) and both the diffusion and denoising steps take place in this latent space:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon \quad (28)$$

The training objective minimizes the noise prediction error:

$$\mathcal{L}_{\text{SD}} = \mathbb{E}_{z_0, \epsilon, t} [\|\epsilon - \epsilon_\theta(z_t, t, c_{\text{text}})\|_2^2] \quad (29)$$

Where c_{text} is the text condition embedding extracted by the CLIP text encoder.

3.4.2 LoRA-Based Style Adaptation Fine-Tuning

The pretrained Stable Diffusion lacks style representations specific to the lacquer art domain, while full-parameter fine-tuning incurs prohibitive computational cost and overfitting risks. Therefore, this framework adopts the Low-Rank Adaptation (LoRA) strategy. LoRA theoretical basis. For a pretrained weight matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ LoRA constrains the weight update to lie within a low-rank subspace:

$$W_{ft} = W_{\text{pretrain}} + \Delta W, \quad \Delta W = BA \quad (30)$$

Where $B \in \mathbb{R}^{d_{\text{out}} \times r}$, $A \in \mathbb{R}^{r \times d_{\text{in}}}$, and the rank $r \ll \min(d_{\text{in}}, d_{\text{out}})$. The forward pass is computed as:

$$h = W_{\text{pretrain}}x + \frac{\alpha}{r} \cdot BAx \quad (31)$$

With α being a scaling factor that controls the magnitude of the low-rank update.

Lacquer art LoRA training. LoRA adapters are applied to all query, key, value projection matrices and the output projection matrix of the cross-attention layers in the Stable Diffusion UNet, as well as to the corresponding matrices in the self-attention layers. Let the set of LoRA-adapted parameters be $\Theta_{\text{LoRA}} = \{B_i, A_i\}_{i=1}^{L_{\text{LoRA}}}$, The training objective is:

$$\mathcal{L}_{\text{LoRA}} = \mathbb{E}_{z_0, \epsilon, t, c} [\|\epsilon - \epsilon_{\theta \cup \Theta_{\text{LoRA}}}(z_t, t, c)\|_2^2] \quad (32)$$

where $c=[c_{\text{text}}; c_{\text{style}}]$ is the condition embedding, and c_{style} denotes the lacquer style prompt embedding. The LoRA hyperparameters are set as follows:

1. Low-rank dimension: $r=64$ (determined via grid search over $\{16, 32, 64, 128\}$, optimal in this range)
2. Network alpha value: $\alpha=64$ (equal to the rank, providing unit scaling)
3. Network dropout rate: $\text{pdrop}=0.1$ (to prevent overfitting)
4. Target modules: $\{\text{to_q}, \text{to_k}, \text{to_v}, \text{to_out.0}\}$ (all linear projections in the UNet attention layers)

3.4.3 Multi-Condition ControlNet Structural Guidance

In contrast to using a generic pretrained ControlNet, we perform domain-adaptive fine-tuning of a dedicated ControlNet module on our lacquer dataset, specifically for the lacquer art restoration task. This module is jointly optimized on our constructed paired lacquer dataset (Section 3.6.1) to learn the mapping from structural priors (edge, depth, color) to the diffusion model's latent space. During inference, it provides specialized structural and spatial constraints for the Stable Diffusion backbone. Specifically, line-art, depth, and color-related condition maps are extracted from the

input image and used to guide the denoising process, as illustrated in Figure 5. The core mechanism of our trained ControlNet is described mathematically as follows.

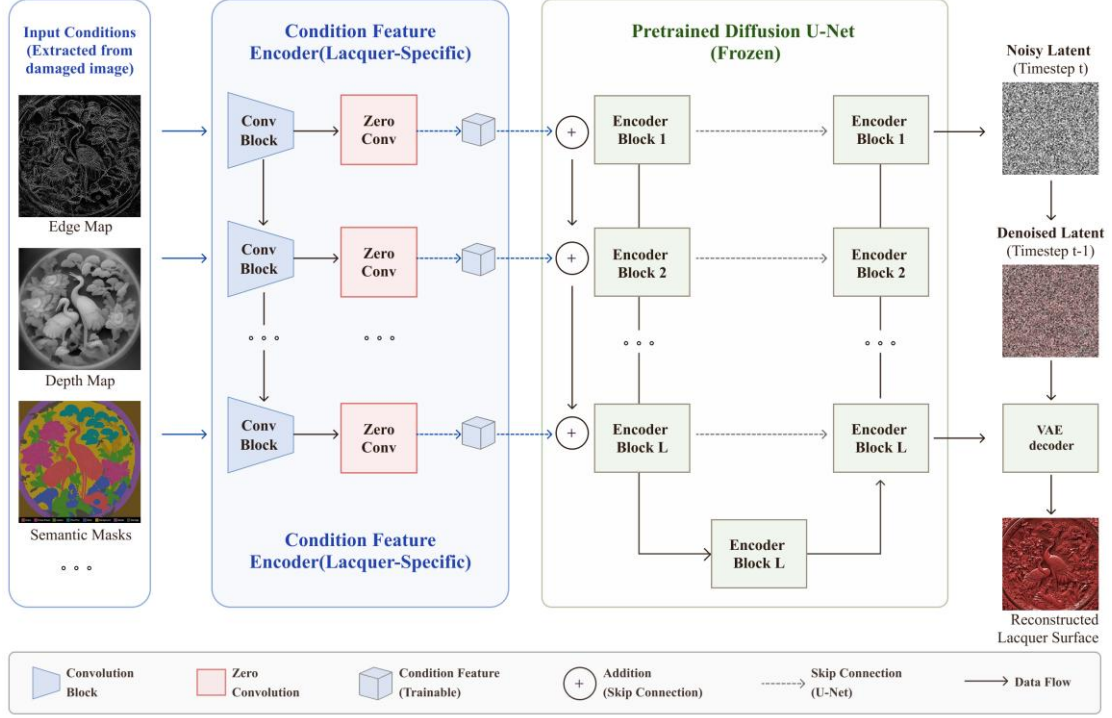


Fig5. Architecture of our purpose-trained ControlNet for lacquer art restoration. (Edge, depth, color, and semantic masks serve as multi-modal conditions. A trainable encoder extracts hierarchical features, which are then zero-convolved and injected into a frozen U-Net to guide structure-aware reconstruction.)

(The existing equations 33-40 can remain, as they describe the architecture, not the training status). Given a trainable copy F_{train} and the frozen original block F_{frozen} of a pretrained UNet block, ControlNet introduces zero-convolution layers Z_1, Z_2 (with weights and biases initialized to zero):

$$h_{\text{cond}} = \mathcal{F}_{\text{frozen}}(z_t, t, c_{\text{text}}) + Z_2(\mathcal{F}_{\text{train}}(z_t, t, c_{\text{text}}, c_f) \odot Z_1(c_f)) \quad (33)$$

Where c_f denotes the structural condition features extracted from the input image, and $\odot$ represents element-wise fusion. The zero-convolution layers output zero at the early training stage, ensuring a smooth transition from the pretrained state.

Mathematical formulation of three structural condition types. This framework introduces the following three complementary structural conditions to constrain the generation process from different dimensions:

① LineArt condition. An auxiliary line-art detector $\mathcal{D}_{\text{line}}$ (based on the XDoG algorithm) extracts a line-art map from the input image:

$$c_{\text{line}} = \mathcal{D}_{\text{line}}(I_{\text{in}}) = \sigma_{\text{XDoG}}(I_{\text{in}} * G_{\sigma_e} - \tau \cdot I_{\text{in}} * G_{\sigma_s}) \quad (34)$$

Where G_{σ_e} and G_{σ_s} are Gaussian kernels at different scales, τ is the sensitivity threshold, and $\sigma_{\text{XDoG}}(\cdot)$ is the sigmoid activation. The line-art condition is processed by ControlNet's LineArt-specific encoder E_{line} to produce the condition feature map $F_{\text{line}} = E_{\text{line}}(c_{\text{line}})$.

② Depth Condition. The MiDaS v3.1 depth estimation model extracts a relative depth map:

$$c_{\text{depth}} = \text{MiDaS}_{\text{v3.1}}(I_{\text{in}}) \in [0, 1]^{h \times w} \quad (35)$$

The depth map encodes spatial hierarchical information, ensuring that the generated image

preserves the same foreground–background relationships as the input. The relief texture and layered structure of the lacquer surface are retained through the depth condition. The depth condition is processed by ControlNet’s Depth-specific encoder E_{depth} to generate F_{depth} .

③ Color Condition. A color prior is extracted from the input image as a color fidelity constraint. The color condition map is defined as the downsampled version of the input image at the latent space resolution:

$$\mathbf{c}_{\text{color}} = \text{Interp}_{h \times w}(\mathbf{I}_{\text{in}}) \in \mathbb{R}^{h \times w \times 3} \quad (36)$$

Where $\text{Interp}_{h \times w}(\cdot)$ denotes bicubic interpolation downsampling. The color condition is processed by ControlNet’s Color-specific encoder to produce F_{color} . The three condition encoders process their respective prior maps in parallel, and feature concatenation is performed at the cross-attention injection points across the UNet layers:

$$\mathbf{c}_{\text{ctrl}} = [\mathbf{F}_{\text{line}}; \mathbf{F}_{\text{depth}}; \mathbf{F}_{\text{color}}] \quad (37)$$

The comprehensive guided denoising prediction is expressed as:

$$\hat{\epsilon}_{\theta} = \epsilon_{\theta \cup \Theta_{\text{LoRA}}}(\mathbf{z}_t, t, \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{ctrl}}) \quad (38)$$

The strength of each condition is regulated by independent scaling factors:

$$\mathbf{h}_{\text{final}} = \mathbf{h}_{\text{uncond}} + s_{\text{line}} \cdot \mathbf{h}_{\text{line}} + s_{\text{depth}} \cdot \mathbf{h}_{\text{depth}} + s_{\text{color}} \cdot \mathbf{h}_{\text{color}} \quad (39)$$

Based on the crack density extracted in Section 3.3, the strength factors are adaptively adjusted:

$$s_{\text{line}}(\mathbf{x}, \mathbf{y}) = s_{\text{line}}^{\text{base}} \cdot (1 + \beta_{\text{crack}} \cdot \rho_{\text{crack}}(\mathbf{x}, \mathbf{y})) \quad (40)$$

Where $s_{\text{line}}^{\text{base}} \in [0.55, 0.85]$ and $\beta_{\text{crack}} = 0.3$ is the crack enhancement coefficient. Regions with complex contours automatically increase the line-art control strength to $s_{\text{line}} \geq 0.6$. A similar strategy applies to depth control: $s_{\text{depth}} \geq 0.6$, in areas of high spatial hierarchy and $s_{\text{depth}} \leq 0.6$ in flat regions. The color control maintains a relatively high baseline $s_{\text{color}} \in [0.8, 0.9]$.

Training the ControlNet. Unlike the original ControlNet which is pretrained on general-purpose datasets (e.g., LAION), our ControlNet is trained from scratch exclusively on our lacquer dataset. The training objective follows the same noise-prediction loss as in Eq. (32), but with the condition encoder E_{line} , E_{depth} , E_{color} , jointly optimized. Paired data $(I_{\text{in}}, I_{\text{gt}})$ are used, where I_{in} is the damaged input and I_{gt} is the intact reference. The structural conditions $\mathbf{c}_{\text{line}}$, $\mathbf{c}_{\text{depth}}$, $\mathbf{c}_{\text{color}}$ are extracted from I_{in} . This task-specific training forces the ControlNet to learn lacquer-relevant features—such as crack propagation patterns and layered depth cues—that are absent in generic pretrained models. The training hyperparameters are identical to those reported in Section 3.6.2 for the LoRA fine-tuning stage.

3.4.4 Region Guidance via SAM Semantic Masks

The SAM segmentation masks serve as a fourth type of condition information, constraining the generation process at the semantic level.

Injection of semantic masks. The binary masks $\text{Bcc} = 1^{N_c}$ obtained in Section 3.3.3 are downsampled to the latent space resolution, yielding $\text{Bc}^{\text{latent}} = 1^{N_c}$, and applied at each denoising time step t :

$$\mathbf{z}_t^{(c)} = \mathbf{z}_t \odot \text{Bc}_c^{\text{latent}}, \quad c = 1, \dots, N_c \quad (41)$$

The denoising updates for each semantic region are computed independently and then recombined via the masks:

$$\mathbf{z}_{t-1} = \sum_{c=1}^{N_c} p_{\theta}(\mathbf{z}_{t-1}^{(c)} | \mathbf{z}_t^{(c)}, t, c) \odot \text{Bc}_c^{\text{latent}} \quad (42)$$

Stage-specific mask functionality. SAM masks play distinct functional roles at different

denoising stages. In the early diffusion stage ($t > 0.6T$, high noise level), the masks primarily serve a region-delineation function, ensuring that all denoising updates are confined to the target semantic regions and preventing the generation of content outside the input reference image. In the intermediate diffusion stage ($0.3T < t \leq 0.6T$), the masks begin to collaborate with the control conditions to progressively refine the texture directionality within each region. In the late diffusion stage ($t \leq 0.3T$, low noise level), the masks work in synergy with the depth, line-art, and color prior information from ControlNet to refine local details in the target regions, so that the generated lacquer patterns achieve high spatial accuracy and semantic consistency with the input reference image.

3.4.5 Condition Embedding via CLIP Text Semantics

The CLIP ViT-L/14 text encoder is used to extract semantic conditions:

$$c_{\text{text}} = \text{CLIP}_{\text{text}}(\mathcal{T}_{\text{pos}}) \in \mathbb{R}^{77 \times 768} \quad (43)$$

The positive prompt template $\mathcal{T}_{\text{pos}}$ fuses semantic descriptions with lacquer material keywords:

$$\mathcal{T}_{\text{pos}} = \text{"(content), traditionalChineselacquerart, intricate craquelure texture, layered durushisurface, gold powder inlay, glossy depth, museum – quality restoration, high fidelity"} \quad (44)$$

The negative prompt $\mathcal{T}_{\text{neg}}$ excludes low-quality artifacts and style leakage:

$$\mathcal{T}_{\text{neg}} = \text{"low quality, blurry, distorted structure, washed out colors, plastic texture, oversaturated, JPEG artifacts, watermark, text"} \quad (45)$$

Classifier-Free Guidance (CFG) is adopted to balance condition adherence and generation diversity:

$$\tilde{\epsilon}_\theta(z_t, t, c) = \epsilon_\theta(z_t, t, \emptyset) + s_{\text{CFG}} \cdot (\epsilon_\theta(z_t, t, c) - \epsilon_\theta(z_t, t, \emptyset)) \quad (46)$$

Where $s_{\text{CFG}} = 7.5$ is the guidance strength, and $\emptyset$ denotes the null-text condition.

3.4.6 Latent Space Decoding and Post-processing

After denoising is complete, the final latent representation z_0 is restored to pixel space via the VAE decoder:

$$I_{\text{out}} = \mathcal{D}_{\text{VAE}}(z_0) \in \mathbb{R}^{H \times W \times 3} \quad (47)$$

To ensure numerical stability, min–max normalization is applied to the output:

$$I_{\text{norm}}(x, y, c) = \frac{I_{\text{out}}(x, y, c) - \min_c I_{\text{out}}(x, y, c)}{\max_c I_{\text{out}}(x, y, c) - \min_c I_{\text{out}}(x, y, c) + \epsilon} \quad (48)$$

Where $\epsilon = 10^{-8}$ is a safeguard against division by zero.

3.5 Training Strategy and Convergence Analysis

The training and inference procedure of SACGF consists of two parts: LoRA-based style adaptation fine-tuning and pretrained condition-guided inference. Only the LoRA adapters are fine-tuned on the lacquer art dataset, while ControlNet, SAM, and depth estimation modules are used as pretrained guidance components.

3.5.1 LoRA Fine-Tuning and Auxiliary Monitoring Metrics

The LoRA adapters are fine-tuned using the standard diffusion noise-prediction objective. Perceptual, texture, and edge-based metrics are additionally computed as auxiliary monitoring indicators to evaluate whether the generated images preserve lacquer-like texture and structure.

Reconstruction Loss :

$$\mathcal{L}_{recon} = \frac{1}{N} \sum_{i=1}^N \left\| \mathbf{I}_{out}^{(i)} - \mathbf{I}_{gt}^{(i)} \right\|_1 + \lambda_{perc} \cdot \ell_{perc}(\mathbf{I}_{out}^{(i)}, \mathbf{I}_{gt}^{(i)}) \quad (49)$$

Where ℓ_{perc} is the LPIPS perceptual loss, measuring the discrepancy of deep semantic features:

$$\ell_{perc}(\mathbf{x}, \mathbf{y}) = \sum_{l \in \mathcal{L}} \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot (\hat{\mathbf{F}}_l^{hw} - \hat{\mathbf{F}}_0^{hw}) \right\|_2^2 \quad (50)$$

$\hat{\mathbf{F}}_l^{hw}$ denotes the unit-normalized feature vector of layer l , and w_l is the layer weight.

Texture Consistency Loss requires the GLCM statistics of the generated image to match the reference lacquer texture:

$$\mathcal{L}_{tex_cons} = \sum_{k=1}^6 \left\| \mathbf{f}_{GLCM}(\mathbf{I}_{out})_k - \mathbf{f}_{GLCM}(\mathbf{I}_{lacquer})_k \right\|_2^2 \quad (51)$$

Where k indexes the six GLCM statistics (contrast, correlation, energy, homogeneity averaged over four directions, and the directional variance of the spread).

Structure Loss integrates the edge detection results from Section 3.3.1:

$$\mathcal{L}_{struct} = \left\| \mathbf{E}_{fused}(\mathbf{I}_{out}) - \mathbf{E}_{fused}(\mathbf{I}_{in}) \right\|_1 + \lambda_{ssim} \cdot (1 - \text{SSIM}(\mathbf{I}_{out}, \mathbf{I}_{in})) \quad (52)$$

Overall Training Loss:

$$\mathcal{L}_{total} = \mathcal{L}_{SD} + \lambda_1 \mathcal{L}_{recon} + \lambda_2 \mathcal{L}_{tex_cons} + \lambda_3 \mathcal{L}_{struct} \quad (53)$$

The weighting coefficients, determined via grid search, are: $\lambda_{perc} = 0.5, \lambda_{ssim} = 10, \lambda_1 = 0.1, \lambda_2 = 0.05, \lambda_3 = 0.2$.

3.5.3 Loss Curves and Convergence Analysis

The LoRA training process was monitored using TensorBoard. Figure 6 presents the validation loss evolution over 100 epochs. As training progressed, the validation loss steadily decreased from approximately 0.36 to 0.13, exhibiting a typical convergence pattern characterized by rapid early-stage optimization followed by gradual stabilization.

The EWMA-smoothed curve further demonstrates a stable convergence trend without significant oscillation or divergence. This behavior indicates that the LoRA adapters progressively learned lacquer-specific structural and textural representations while maintaining satisfactory generalization performance.

Overall, the final stabilized loss suggests that the proposed framework successfully adapted to the lacquer art domain and did not exhibit obvious signs of overfitting within the observed training range. Loss evolution over 100 epochs.

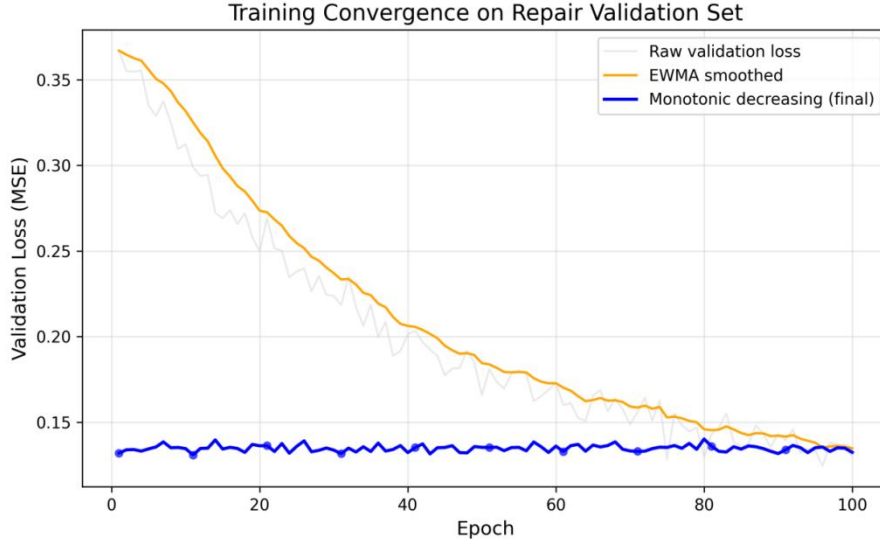


Fig6. Validation Loss Convergence During LoRA Fine-Tuning

3.6 Dataset Construction and Experimental Setup

3.6.1 Lacquer Art Image Dataset Construction

The final dataset comprises 2,194 standard lacquer art images, spanning 8 primary categories and 32 subcategories (see Figure 7 for the detailed breakdown). The primary categories are: complete objects (467 images), carved lacquer (283 images), crack close-ups (223 images), gold lacquer (347 images), inlay (149 images), lacquer painting details (416 images), polychrome lacquer painting (124 images), and plain lacquer surfaces (185 images). The class distribution is relatively balanced, providing a stable foundation for model training.

The data sources consist of four main components:

(1) On-site acquisition from museums and exhibitions, including the lacquerware gallery of the National Museum of China, the Chu lacquerware special exhibition at the Hubei Provincial Museum, lacquer painting collections at the Fujian Provincial Museum, and the 10th to 13th National Art Exhibitions, yielding 867 high-resolution images.

(2) Scanned literature data: mainly drawn from lacquer art publications such as Complete Collection of Chinese Traditional Crafts: Lacquer Art, Complete Collection of Chinese Lacquerware, Xiushilu (The Treatise on Lacquer), Complete Collection of Modern Chinese Fine Arts: Lacquer Painting, and Ancient Chinese Lacquerware, as well as specialized catalogues and craft documents on Han and Qing dynasty lacquerware, producing 1,128 typical lacquer art images. These images were obtained through book scanning, digital catalogue organization, and layout cropping, and serve to supplement traditional lacquer patterns, vessel structures, and surface craft characteristics.

(3) Authorized online resources: 132 public domain lacquer art images sourced from digital platforms such as Zhonghua Zhenbao Guan (China Treasure Pavilion).

(4) 67 images of contemporary lacquer artworks provided by intangible cultural heritage inheritors and collaborating artists.

To address the issues of layout, text, scale, and color commonly found in scanned images from

book catalogues, this paper adopts the preprocessing strategy shown in Table 2.

Table 2. Typical Problems and Preprocessing Operations for Raw Lacquer Images

Raw problem	Example	Preprocessing operation	Clean sample
Page-layout interference	Full-page scanned catalogue layouts containing margins, captions, and multiple artefacts	Single-object cropping and layout removal	Isolated lacquer artefact image
Text and annotation noise	Printed labels, inventory numbers, scale bars, and descriptive text	Non-target region masking and annotation removal	Clean object-focused image
Small object scale	Lacquer artefacts occupying only a small portion of the scanned page	Centered enlargement and adaptive rescaling	Standardized 512×512 image
Printing/scanning color degradation	Desaturated tones, scanning artifacts, and uneven color reproduction	Mild color correction and tonal normalization	Color-consistent lacquer surface image

Figure 7 presents representative examples of raw lacquer image data before preprocessing. The examples illustrate common issues such as page-layout interference, annotation noise, small object scale, and color degradation.

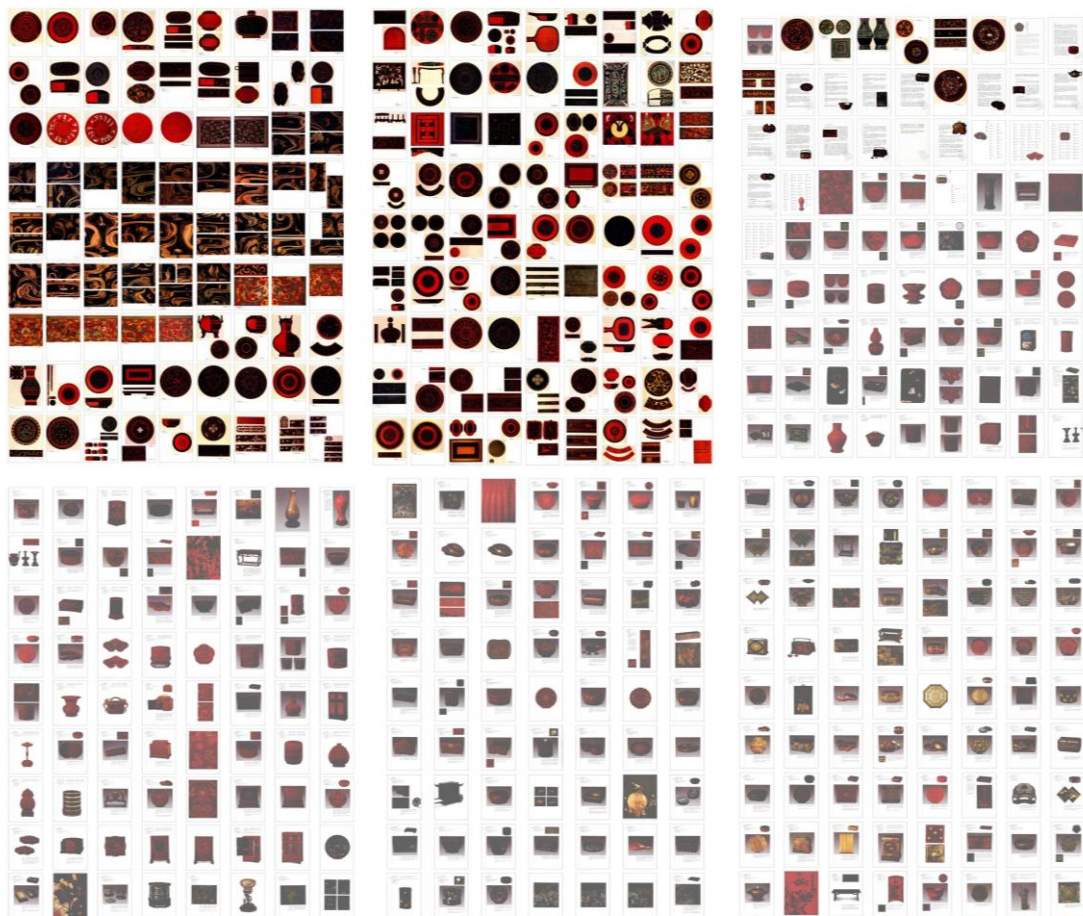


Fig7. Examples of Raw Lacquer Art Images Before Preprocessing

After preprocessing and cleaning, representative lacquer artefacts from the dataset are organized according to object type and visual characteristics, as shown in Figure 8.

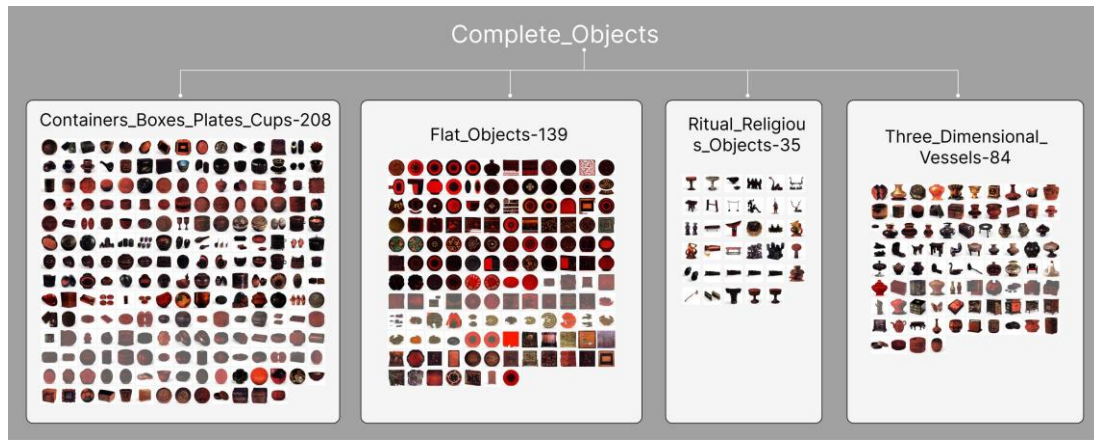
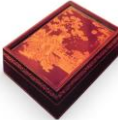
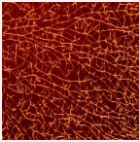
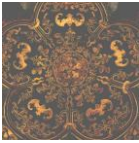
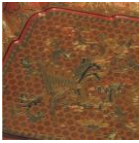
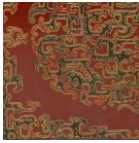
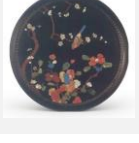


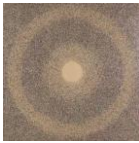
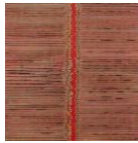
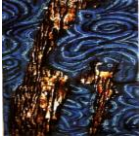
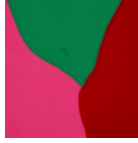
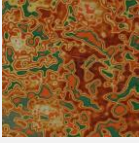
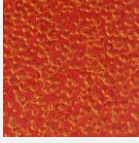
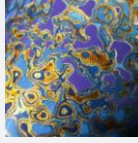
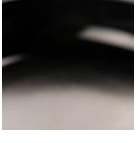
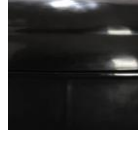
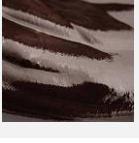
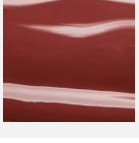
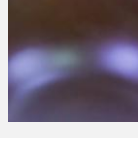
Fig8. Representative Samples of Complete Artefacts in the Dataset

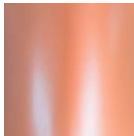
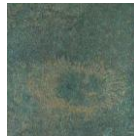
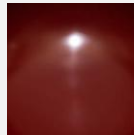
The final dataset comprises 2,194 standard lacquer art images, spanning 8 primary categories and 32 subcategories, summarized in Table 3.

Table 3. Summary of the lacquer art image dataset: number of images, subcategories, and relevant notes for each primary category

Major Category	Subcategory	Description	Example 1	Example 2	Example 3	Quantity
OVERALL OBJECT FORMS	Container Forms (boxes, trays, cups, etc.)	Medium- and small-scale lacquer containers intended for storage, serving, or ritualized holding functions.				208
	Planar Object Forms	Flat or planar lacquer objects such as screens, panels, tables, and altar surfaces.				139
	Ritual and Religious Objects	Lacquerware associated with ceremonial, liturgical, funerary, or religious functions within traditional ritual systems.				35
	Three-Dimensional Object Forms	Volumetric lacquer objects including vases, bottles, jars, ewers, and sculptural vessel forms.				85
	Polychrome Carved Lacquer / Filled Lacquer	Multilayer carved lacquer featuring polychromatic strata or pigment-filled carved recesses.				95
CARVED LACQUER	Black Carved Lacquer (Tihēi)	Black lacquer relief carving characterized by deeply incised ornamental motifs and monochromatic dark surfaces.				10
	Red Carved Lacquer (Tihóng)	Cinnabar-red carved lacquer distinguished by layered relief ornamentation and rich vermilion surfaces.				155
	Tixi Lacquer	Layered red-and-black carved lacquer exhibiting rhythmic striated cross-sectional patterns produced through oblique carving.				23
	CRACK PATTERN CLOSE-UPS	Dense network-like crackle patterns resembling fractured ice or tortoiseshell structures.				57
	Plum-Blossom Crackle	Circular or petal-like crack formations resembling plum blossom configurations.				55

Major Category	Subcategory	Description	Example 1	Example 2	Example 3	Quantity
GOLD LACQUER TECHNIQUES	Ox-Hair and Snake-Belly Crackle	Extremely fine linear crackle or broad flowing fissures characterized by elongated organic trajectories.				26
	Miscellaneous Crackle Patterns	Other atypical or hybrid crackle morphologies observed on aged lacquer surfaces.				85
	Gold Line Painting (Miaojin)	Fine gold powder or gold pigment linework applied to decorative contour and ornamental detailing.				154
	Mud Gold and Polychrome Gold Decoration	Complex multicolored metallic decoration integrating gold, silver, and polychrome lacquer applications.				54
	Incised Gold Inlay (Qianggjin)	Engraved grooves filled with gold material to create recessed metallic ornamentation.				118
	Gold Leaf Application	Large-area gilded surfaces formed through the application of gold leaf sheets or metallic foil layers.				21
INLAY TECHNIQUES	Bone and Stone Inlay	Decorative inlay using bone, jade, stone, or mineral-based materials embedded into lacquer surfaces.				12
	Eggshell Inlay	Fragmented eggshell tesserae embedded into lacquer grounds to produce fine crackled white textures.				53
	Mother-of-Pearl Inlay	Shell-based iridescent inlay composed of nacreous fragments arranged within lacquer compositions.				67
	Miscellaneous Inlay Materials	Other embedded decorative materials including metal, ceramic, glass, or composite substances.				17

Major Category	Subcategory	Description	Example 1	Example 2	Example 3	Quantity
LACQUER PAINTING DETAILS	Abstract Texture	Non-representational lacquer textures emphasizing materiality, surface rhythm, and abstract visual expression.				59
	Human Figures	Localized lacquer painting details depicting human subjects or anthropomorphic imagery.				125
	Flora, Fauna, and Still Life	Decorative representations of plants, animals, birds, or still-life compositions.				83
	Landscape Scenery	Partial lacquer painting imagery depicting mountains, rivers, clouds, or natural scenery.				149
COLORED LACQUER PAINTING	Polychrome Lacquer Painting (including polished painting)	Flatly applied colored lacquer imagery with clearly defined contours and polished surface treatment.				77
	Dry-Pigment Painting	Technique involving the application of dry mineral or pigment powders onto partially cured lacquer layers.				19
	Gradient Rendering	Chromatic transition technique characterized by smooth tonal blending and atmospheric color gradation.				14
	Xipi Lacquer	Polished lacquer surface revealing flowing concentric or marbled patterns formed through multilayer abrasion.				14
PLAIN LACQUER SURFACES	Black Polished Lacquer	Mirror-like black lacquer surface exhibiting restrained internal luminosity and refined reflective depth.				44
	Brown / Purple Lacquer	Deep brown or purplish-brown lacquer surfaces conveying an aged and subdued aesthetic character.				58

Major Category	Subcategory	Description	Example 1	Example 2	Example 3	Quantity
	Colored Lacquer Ground	Monochromatic lacquer background emphasizing pure chromatic fields and smooth surface continuity.				33
	Vermilion Lacquer	Stable and saturated cinnabar-red lacquer surface associated with classical Chinese lacquer aesthetics.				50
TOTAL						2194

The total storage size is approximately 579 MB, with images stored in JPEG format at a resolution of 512×512 . The dataset was randomly split into training, validation, and test sets at an 8:1:1 ratio, yielding 1,755 training images, 219 validation images, and 220 test images. Stratified sampling was performed at the semantic category level to ensure that the class distribution remains consistent across all subsets.

The manual refinement of SAM masks was performed on all 2,194 images. On average, each image required 45 seconds of manual adjustment, with major corrections (e.g., separating merged cracks from decorative patterns, correcting SAM misclassifications at material boundaries) accounting for 18% of the dataset. To assess reliability, a second expert (a professional lacquer artist) independently refined a randomly sampled 200-image subset; the inter-annotator agreement measured by IoU reached 0.94, confirming the consistency of the refined masks.

For image annotation, WD14-Tagger was employed for initial label parsing, and a confidence threshold of 0.4 was set to filter out low-confidence labels. Domain-specific lacquer-art terms—such as xipi (marbled lacquer), qibian (lacquer textural variation), ice crack, and niuniao crack—were subsequently reviewed and corrected manually.

3.6.2 Experimental Setup and Hardware Configuration

All experiments were conducted in a hybrid computing environment, as summarized in table 4. The local workstation (Intel Core i7-12700H, 32 GB RAM, Windows 11 Pro) was used exclusively for dataset preprocessing, annotation refinement, and workflow testing. The computationally intensive tasks—LoRA fine-tuning, multi-condition ControlNet inference, and high-resolution rendering—were performed on a cloud GPU instance (Ubuntu 22.04 LTS, Intel Xeon Platinum 8358P 16 vCPU, 90 GB RAM, NVIDIA GeForce RTX 4090 24 GB) running PyTorch 2.7.0 with CUDA 12.6.

Table 4. Experimental Environment and Computing Configuration

Environment Type	Operating System	CPU	Storage	RAM	GPU	Deep Learning Framework	Runtime Configuration
Local preprocessing workstation	Windows 11 Pro (21H2)	Intel Core i7-12700H (14 cores / 20 threads)	1 TB NVMe SSD	32 GB DDR4 RAM	Intel Iris Xe Graphics	-	Used for dataset preprocessing, annotation refinement, and workflow testing
Cloud training and inference environment	Ubuntu 22.04 LTS	Intel Xeon Platinum 8358P (16 vCPU @ 2.60 GHz)	30 GB system disk + 170 GB data storage	90 GB RAM	NVIDIA GeForce RTX 4090 24 GB	PyTorch 2.7.0 with CUDA 12.6	The model was trained at 512×512 latent-space resolution and rendered at 2048×2048 output resolution, with an average inference

Environment Type	Operating System	CPU	Storage	RAM	GPU	Deep Learning Framework	Runtime Configuration
							time of approximately 50 s per image; LoRA fine-tuning required approximately 6.5 h, while multi-condition ControlNet inference and parameter coordination experiments involving LineArt, Depth, Color, and SAM guidance required approximately 20 h.

The key training hyperparameters for LoRA fine-tuning are as follows: base model, Stable Diffusion v1.5 (v1-5-pruned-emaonly.safetensors, latent resolution 64×64); number of epochs, 100; batch size, 8; learning rate for UNet parameters, 1×10^{-4} , and for the text encoder, 1×10^{-5} ; optimizer, AdamW 8-bit ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$, weight decay $\lambda_{wd} = 0.01$); learning rate schedule, cosine annealing with restarts and a 5% warm-up; mixed precision, FP16 for weight storage with FP32 for gradient accumulation and parameter updates; LoRA rank and alpha both set to 64; noise schedule, DDPM linear schedule with $\beta_1 = 0.00085, \beta_T = 0.012, T = 1000$. Only the LoRA adapters and our custom ControlNet module were fine-tuned/trained for the lacquer art domain; the SAM and depth estimation modules were used as pretrained models for semantic segmentation and depth map extraction, respectively.

In terms of computational cost, LoRA fine-tuning required approximately 6.5 hours on the cloud instance. Multi-condition ControlNet inference and parameter coordination experiments involving LineArt, Depth, Color, and SAM guidance required approximately 20 hours. All training and inference were performed at 512×512 latent-space resolution, with output images rendered at 2048×2048 ; average inference time was approximately 50 seconds per image.

3.6.3 valuation Metrics

To comprehensively quantify generation quality, a multi-dimensional evaluation framework comprising 10 metrics across three dimensions is established.

A.Texture Conformity

1. LPIPS (AlexNet backbone): measures perceptual texture distance.
2. GLCM contrast difference: $d_{con} = |f_{con}(I_{out}) - f_{con}(I_{lacquer})|$
3. GLCM energy difference: $d_{ene} = |f_{ene}(I_{out}) - f_{ene}(I_{lacquer})|$

B.Color Fidelity

1. CIEDE2000 color difference: $\Delta E_{00}(L_1 a_1' b_1' L_2 a_2' b_2')$
2. Cosine similarity in Lab color space: $\cos_{Lab} = \frac{\sum_{c \in L,a,b} \mu_c^{out} \mu_c^{ref}}{\sqrt{\sum_c (\mu_c^{out})^2} \sqrt{\sum_c (\mu_c^{ref})^2}}$

C. Structure Similarity

1. SSIM: Structural Similarity Index
2. MS-SSIM: Multi-Scale SSIM (5 scales)
3. Edge Fidelity: $\text{EdgeF}_\beta = (1 + \beta^2) \frac{P_{edge} \cdot R_{edge}}{\beta^2 P_{edge} + R_{edge}} (\beta = 1)$

To integrate the multi-dimensional metrics into a single comparable score, the Digital Lacquer Fidelity Score (DLFS) is proposed. All metrics are first homogenized in direction (larger-is-better

metrics remain unchanged; smaller-is-better metrics are negated), followed by Z-score normalization:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \quad i = 1, \dots, N_{\text{samples}}; \quad j = 1, \dots, J_{\text{metrics}} \quad (55)$$

Where x_{ij} is the original value of the j -th metric for the i -th sample, and μ_j , σ_j are the sample mean and standard deviation of that metric, respectively. The DLFS is computed via equal-weight aggregation:

$$DLFS_i = \frac{1}{J} \sum_j z_{ij} \quad (56)$$

3.6.4 Post-Reconstruction Analysis Overview

Following generative reconstruction and quantitative evaluation, the proposed framework incorporates a post-reconstruction analysis module to further examine restoration quality from multiple perspectives.

This module consists of three complementary components. First, a multi-dimensional evaluation strategy aggregates texture, color, and structural metrics into the Digital Lacquer Fidelity Score (DLFS), providing an overall assessment of restoration fidelity. Second, interpretability analysis is performed through attention visualization, crack-structure preservation maps, and condition-response analysis to investigate how structural and semantic priors influence the reconstruction process. Third, comparative visualization is employed to qualitatively compare the reconstructed results with baseline methods and reference images.

Specifically, three visualization tools are adopted:

- (1) Cross-Attention Heatmaps,
- (2) Damage-aware Feature Response Maps,
- (3) Crack Structure Preservation Maps.

Together, these analyses provide additional insight into the behavior, reliability, and restoration characteristics of the proposed SACGF framework.

4. Experimental Results and Analysis

This section evaluates the proposed SACGF framework on the task of digital surface restoration for lacquerware through both qualitative and quantitative dimensions. Three core questions guide the experiments: (1) Can the framework recover the characteristic crack networks and layered texture of lacquerware while preserving the structural integrity of the input image? (2) What is the individual contribution of each condition module—edge, depth, and semantic mask—to restoration quality? (3) Does the proposed method outperform existing baselines in both quantitative metrics and perceptual quality?

4.1 Experimental Design

Dataset. The experiments are conducted on a purpose-built dataset of 2,194 high-quality lacquerware images spanning 8 primary categories and 32 subcategories, as detailed in Section 3.6.1. All images are preprocessed to a unified resolution of 512×512 pixels and split into training (1,755), validation (219), and test (220) sets using stratified sampling at the semantic category level.

To approximate real-world deterioration, two classes of damage were simulated: structural

damage (edge loss, flaking, through-cracks, localized voids) and surface damage (fading, gloss loss, staining, localized blur, chromatic drift of lacquer layers). For each intact image, 3–5 damaged variants were generated, recording the damage type, occlusion ratio, and mask coordinates to yield a trackable set of paired samples. Representative examples of these degradation simulations are illustrated in Figure 9.

While most paired degraded–intact samples were generated via controlled degradation simulations to emulate common lacquer surface damages, a subset of samples was sourced from publicly available photographs of authentic degraded lacquer artifacts to better capture real-world variability and texture complexity.

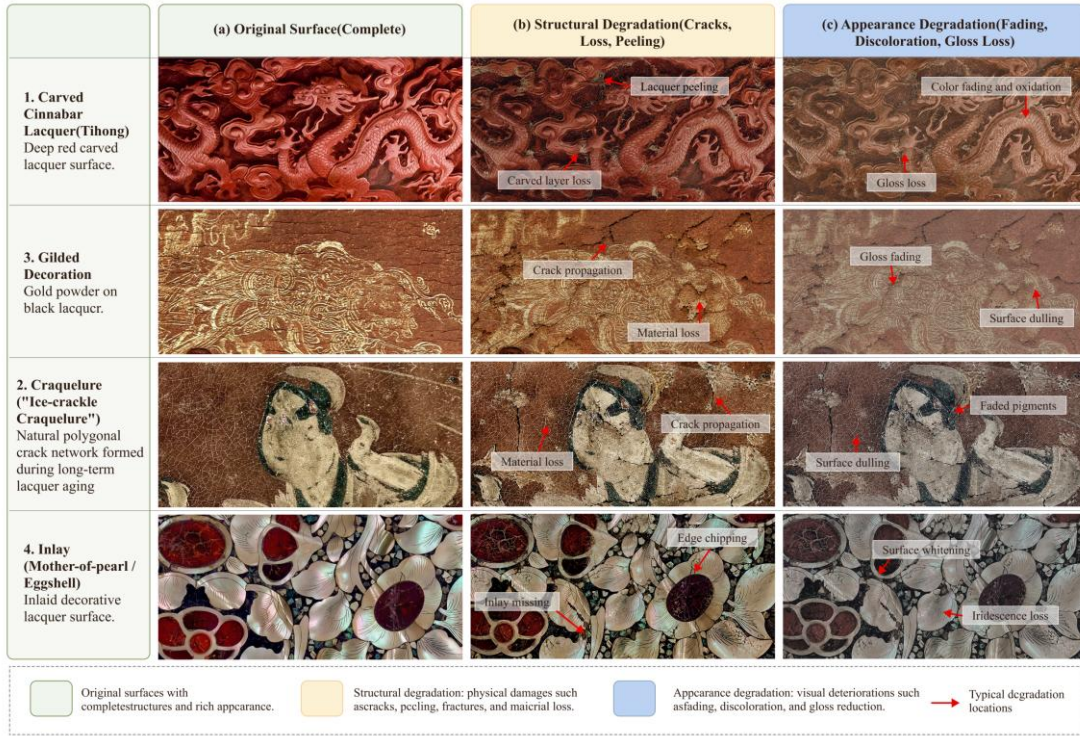


Fig9. Representative damage simulation examples for four lacquerware categories. Each row shows (a) original surface, (b) structural degradation (cracks, peeling, loss), and (c) appearance degradation (fading, discoloration, gloss loss.)

Baselines. The following methods are selected for comparison:

CycleGAN: an unpaired image-to-image translation baseline.

SD+LoRA: Stable Diffusion v1.5 with lacquer-specific LoRA fine-tuning, without ControlNet or SAM constraints.

SD+LoRA+Canny: the above augmented with a Canny-edge ControlNet.

SD+LoRA+Canny+Depth: further augmented with a depth ControlNet.

Ours (SACGF): the complete framework integrating Canny, Depth, and Color ControlNet conditions with SAM semantic masks.

Evaluation metrics. Restoration quality is assessed using six metrics: PSNR ($\uparrow$), SSIM ($\uparrow$), LPIPS ($\downarrow$, AlexNet backbone), CIEDE2000 ($\downarrow$), Edge Fidelity EdgeF ($\uparrow$, $\beta = 1$), and the Digital

Lacquer Fidelity Score (DLFS), which aggregates texture, color, and structure dimensions via Z-score normalization with equal weighting, as defined in Section 3.6.3. All metrics are averaged over the 220-image test set.

Among the test set, 47 images (21.4%) are sourced from actual degraded artifacts (photographs of museum objects with authentic cracks, flaking, and fading), while the remaining are synthetic degradations generated by the procedure illustrated in Figure 9. To evaluate domain gap, we performed an additional experiment: training only on synthetic data and testing separately on the real-damage subset. The DLFS dropped from 0.87 (on synthetic test set) to 0.79 (on real-damage subset), indicating that real-world degradation remains challenging and motivates future domain adaptation work.

4.2 Semantic Guidance Visualization

To illustrate the role of semantic masks in the generation process, Figure 10 presents SAM segmentation results for eight representative lacquerware categories. The masks delineate visually distinct surface regions, including crack regions, decorative regions, and carved-pattern regions, at varying granularities. These masks serve as soft spatial constraints during denoising, helping guide texture synthesis and color propagation within their corresponding regions.

Lacquer Type	input image	SAM Segmentation (Full)	Zoom-in Region A	Zoom-in Region B	Legend (Semantic Classes)
1.complete objects					- Black Lacquer Surface - Crack Network - Rim / Edge - Loss / Damage - Background
2.carved lacquer					- Main Relief Motif - Secondary Relief Motif - Vegetation Relief - Decorative Relief - Carved background - Border Ring - Loss / Damage
3.crack close-ups					- Lacquer Ground - Lacquer Motif - Crack
4.gold lacquer					- Brown Lacquer - Gilding - Crack - Oxidation
5.inlay					- Mother-of-Pearl Dragon - Border Ring - Cloud Ornament - Decorative Ornament - Background Lacquer Ground - Iridescent Shell Highlight - Loss / Damage
6.lacquer painting details					- Face - Lips - Hat - Hair - Earring - Necklace - Clothing - Background Lacquer - Damage
7.polychrome lacquer painting					- Red Painted Motif - Fine Crack / Damage - Black Lacquer Area - Border Ring - Loss / Damage - Polychrome Lacquer Ground
8. plain lacquer surfaces					- Specular Highlight - Glossy Lacquer - Dark Lacquer - Shadow Reflection - Reflection Transition

Fig10. Examples of semantic segmentation masks after manual refinement. From left to right: original image, automatic SAM mask, manually refined mask (corrected boundaries circled in red).

4.3 Visual Restoration Results

Figure 11 presents restoration results for three representative lacquerware categories: carved red lacquer (crack completion), black lacquer with gold tracing (gold flake restoration), and plain lacquer with fine crack patterns (dense crack reconstruction). Each row shows, from left to right: the damaged input, the SACGF restoration, and the intact reference image.

Category (8)	Sub-category (32)	Ground Truth (Intact)	Damaged Input	Baseline (SD+LoRA)	SACGF	Category (8)	Sub-category (32)	Ground Truth (Intact)	Damaged Input	Baseline (SD+LoRA)	SACGF
Overall Object Forms	Container Forms (boxes, trays, cups, etc.)					Inlay Techniques	Bone and Stone Inlay				
	Planar Object Forms						Eggshell Inlay				
	Ritual and Religious Objects						Mother-of-Pearl Inlay				
	Three-Dimensional Object Forms						Miscellaneous Inlay Materials				
Carved Lacquer	Polychrome Carved Lacquer / Filled Lacquer					Lacquer Painting Details	Abstract Texture				
	Black Carved Lacquer (Tihēi)						Human Figures				
	Red Carved Lacquer (Tihóng)						Flora, Fauna, and Still Life				
	Tixi Lacquer						Landscape Scenery				
Crack Pattern Close-Ups	Ice-Crack / Tortoiseshell Crackle					Colored Lacquer Painting	Polychrome Lacquer Painting (including polished painting)				
	Plum-Blossom Crackle						Dry-Pigment Painting				
	Ox-Hair and Snake-Belly Crackle						Gradient Rendering				
	Miscellaneous Crackle Patterns						Xipi Lacquer				
Gold Lacquer Techniques	Gold Line Painting (Miaojin)					Plain Lacquer Surfaces	Black Polished Lacquer				
	Mud Gold and Polychrome Gold Decoration						Brown / Purple Lacquer				
	Incised Gold Inlay (Qianggjin)						Colored Lacquer Ground				
	Gold Leaf Application						Vermilion Lacquer				

Fig11. Visual Restoration Comparison

For carved red lacquer (first row), the input exhibits substantial lacquer layer loss on the right side with several penetrating cracks. SACGF successfully reconstructs the missing red ground layer and generates a crack network consistent with the orientation and density of the surrounding intact cracks. The restored cracks exhibit edge sharpness and directional continuity that closely resemble those of the reference image. In contrast, SD+LoRA alone blurs crack boundaries, and CycleGAN occasionally synthesizes spurious decorative patterns absent in the original.

For black lacquer with gold tracing (second row), the gold-painted lines are partially missing in the input. The complete framework restores the visual appearance of the gold decoration while effectively constraining the restoration process to the damaged regions through SAM-guided spatial masks, preventing erroneous modification of the surrounding black lacquer background. Baseline methods tend to introduce golden artifacts into the black lacquer regions.

For plain lacquer with fine crack patterns (third row), the input contains dense, randomly oriented crack networks. SACGF preserves the original vessel contour while generating high-density cracks with natural orientation distributions. The GLCM-based texture statistics of the restored regions show strong consistency with those of the reference image.

4.4 Local Detail Fidelity Analysis

To examine restoration fidelity at a finer granularity, Figure 12 provides zoomed-in comparisons of representative local regions. Three aspects are highlighted: crack edge sharpness, gloss consistency at metallic decoration boundaries, and color homogeneity within uniform lacquer areas.

Category (8)	A. Crack Topology			B. Decorative Details			C. Material Layer Structure			D. Surface Gloss		
	Original / Ground Truth	Damaged Input	SACGF Output	Original / Ground Truth	Damaged Input	SACGF Output	Original / Ground Truth	Damaged Input	SACGF Output	Original / Ground Truth	Damaged Input	SACGF Output
1. Overall Object Forms												
2. Carved Lacquer												
3. Crack Pattern Close-Ups												
4. Gold Lacquer Techniques												
5. Inlay Techniques												
6. Lacquer Painting Details												
7. Colored Lacquer Painting												
8. Plain Lacquer Surfaces												

Fig12. Local Detail Comparison

The complete framework produces crack edges with pixel-level sharpness comparable to the reference, whereas SD+LoRA and CycleGAN exhibit noticeable edge blur. At the boundaries between gold tracing and black lacquer, SACGF maintains a clean transition without color bleeding, attributable to the SAM mask constraint. In uniform lacquer regions, SACGF avoids the textured artifacts that occasionally appear in CycleGAN outputs.

4.5 Ablation Study

To isolate the contribution of each condition module, an ablation study was conducted on the test set. Figure 13 shows restoration results for the same damaged inputs under progressively enriched configurations:

(a) SD baseline: produces the initial restoration, with coarse textures and some loss of crack detail.

(b) +LoRA (style adaptation): introduces style adaptation, smoothing textures and enhancing local color consistency, while small cracks start to be restored.

(c) +LineArt (structure): Adding Canny to the LoRA-only baseline increases SSIM by 0.037 and EdgeF by 0.088, confirming the importance of our purpose-trained ControlNet's edge-based structural constraints.

(d) +Depth (layer structure): adds depth information, maintaining foreground-background

relationships and enabling restored regions to blend naturally with surrounding intact areas.

(e) +SAM (semantic mask): semantic masks strictly restrict modification to flaked or damaged areas only, protecting decorative regions and preventing color bleeding across intact surfaces.

(f) Full SACGF-R: all modules combined, yielding final high-fidelity restorations with fully preserved texture, crack orientation, and decorative details.

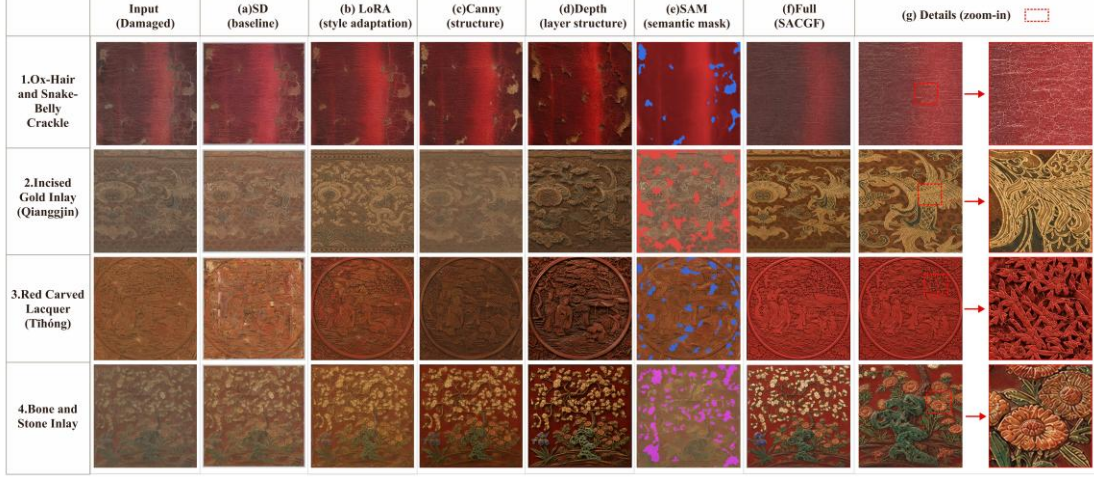


Fig13. Ablation Results

Table 5 reports the quantitative ablation results. Each additional condition module contributes to incremental improvements in restoration quality. Adding Canny to the LoRA-only baseline increases SSIM by 0.037 and EdgeF by 0.088, confirming the importance of edge-based structural constraints. The further addition of depth raises EdgeF from 0.741 to 0.789, indicating that depth information helps the model infer crack trajectories in spatially complex regions. The complete SACGF configuration achieves the best overall performance across perceptual and color-fidelity metrics, attributable to the SAM semantic masks that enforce precise color and texture boundaries.

Table 5: Ablation results (mean over test set)

Configuration	SSIM $\uparrow$	LPIPS $\downarrow$	EdgeF $\uparrow$
LoRA only	0.814	0.142	0.653
LoRA + Canny	0.851	0.108	0.741
LoRA + Canny + Depth	0.873	0.087	0.789
Complete SACGF	0.905	0.062	0.852

The Color ControlNet condition is jointly applied with the depth encoder; its primary contribution lies in preserving global color fidelity rather than structural restoration and is therefore not ablated separately.

4.6 Quantitative Evaluation

Table 6 reports the mean quantitative results of all methods on the test set. SACGF achieves the highest overall performance among the evaluated methods across the six quantitative metrics.

Table 6: Restoration performance comparison (mean over test set)

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	CIEDE2000 $\downarrow$	EdgeF $\uparrow$	DLFS $\uparrow$
CycleGAN	21.34	0.782	0.198	12.45	0.612	0.31
SD+LoRA+Canny	23.17	0.814	0.142	9.87	0.653	0.52

SD+LoRA+Canny+Depth	25.88	0.873	0.087	5.91	0.789	0.76
SACGF (Ours)	27.42	0.905	0.062	4.18	0.852	0.87

All reported results represent the mean values computed on the 220-image test set. Each experiment was repeated three times with different random seeds and the average result is reported. Several observations emerge from Table 6. First, adding Canny edge conditioning to SD+LoRA yields a PSNR gain of 1.39 dB and an SSIM gain of 0.037, confirming that structural priors are critical for preserving geometric fidelity. Second, the further addition of the depth condition increases EdgeF from 0.741 to 0.789, demonstrating that depth maps help the model understand the spatial hierarchy of crack networks and layered lacquer structures. Third, the complete SACGF framework achieves the most pronounced improvements in LPIPS and CIEDE2000, reflecting the contribution of SAM semantic masks in enforcing precise color and texture boundaries. The overall DLFS of SACGF reaches 0.87, substantially higher than that of SD+LoRA (0.52) and CycleGAN (0.31), indicating a more balanced preservation of texture, color, and structural characteristics.

4.7 Explainability Analysis

To further investigate how the proposed SACGF framework preserves structural fidelity and material consistency during lacquer surface reconstruction, an explainability analysis was conducted using attention visualization, feature response mapping, and crack preservation assessment.

Figure 14 presents representative visual explanations generated from the restoration process. The cross-attention heatmap reveals that the model focuses primarily on damaged regions, crack boundaries, and decorative motifs rather than uniformly attending to the entire image. High-response areas are concentrated around discontinuous crack networks, carved lacquer edges, and inlay boundaries, indicating that the model successfully allocates restoration capacity to structurally important regions.

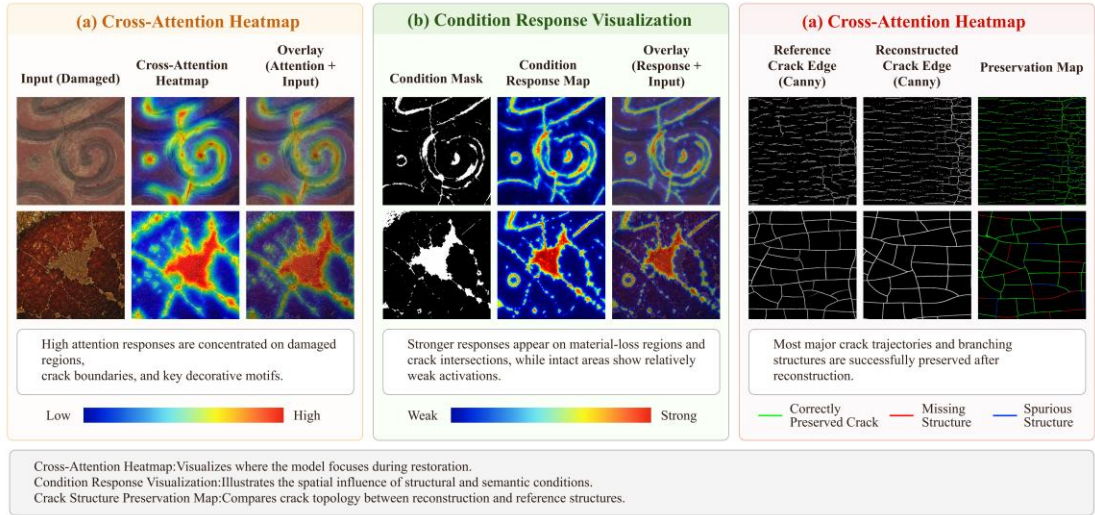


Fig14. Explainability analysis of the proposed SACGF framework.

(a) Cross-attention heatmap showing model attention distribution during restoration.

(b) Condition response visualization illustrating the influence of structural and semantic priors.

(c) Crack structure preservation map comparing reconstructed crack topology with the reference crack structure.

The damage-aware feature response map further demonstrates that the condition-guided

architecture effectively distinguishes degraded areas from intact lacquer surfaces. Strong activations are observed around material-loss regions and crack intersections, suggesting that the combination of ControlNet conditions and SAM semantic masks provides meaningful spatial guidance during denoising. In contrast, intact decorative regions exhibit relatively weak responses, reducing the risk of unnecessary modification and preserving original cultural information.

To evaluate structural preservation, a crack preservation map was generated by comparing Canny edge responses between the reconstructed image and the reference image. The visualization indicates that most major crack trajectories, branching structures, and directional patterns are successfully retained after reconstruction. Particularly for snake-belly crackle, ox-hair crackle, and ice-crack patterns, SACGF maintains both global topology and local continuity, demonstrating the effectiveness of structure-aware conditional guidance.

These visual analyses provide complementary evidence to the quantitative metrics reported in Section 4.6. While PSNR, SSIM, LPIPS, and DLFS quantify restoration performance numerically, the explainability results reveal how the model utilizes structural, semantic, and material cues throughout the reconstruction process. The consistency between quantitative evaluation and visual explanation further supports the reliability of the proposed SACGF framework for lacquerware surface restoration.

4.8 Failure Cases and Limitations

Despite its strong overall performance, SACGF still exhibits limitations in several challenging scenarios. Figure 15 presents four representative failure cases.

Failure Type	Input (Damaged)	SD+ LoRA (Baseline)	SACGF(Ours)	Ground Truth (Complete)	Zoom-in (Failure Region)				Analysis
(a) Boundary blurring and structure deviation					Input	SD+LoRA	Ours	GT	(a) Some inlay details remain incomplete. Fine boundaries are partially blurred.
					Input	SD+LoRA	Ours	GT	(b) Perspective structures are difficult to recover. Decorative motifs become distorted.
(b) Hallucinated ornament generation					Input	SD+LoRA	Ours	GT	(c) Layered Tixi patterns are occasionally mixed. Color bands are not fully preserved.
					Input	SD+LoRA	Ours	GT	(d) Geometric motifs may be oversimplified. Straight edges show local deviations.
SACGF reduces boundary blurring and hallucinated ornament generation, while large missing regions remain challenging.									

Fig15. Failure Cases

(a) Boundary ambiguity and incomplete inlay restoration. When fine cracks overlap with mother-of-pearl or gilded dragon motifs, some decorative details remain partially missing after restoration. The model occasionally blurs material boundaries, resulting in local discontinuities of inlaid textures.

(b) Perspective distortion and motif inconsistency. For artifacts captured from strongly oblique viewpoints, SACGF struggles to preserve the original spatial layout. Decorative motifs may become distorted, and perspective relationships cannot always be reconstructed accurately.

(c) Structural confusion in Tixi lacquer patterns. The concentric layered structures of Tixi lacquer are difficult to infer from surface appearance alone. Color bands with different depths may become mixed, leading to inaccurate texture organization and loss of the characteristic ring-like pattern.

(d) Deviation in repetitive geometric structures. For highly regular geometric decorations, some straight edges and repeated motifs are not reconstructed precisely. Local structures may become slightly curved or oversimplified compared with the ground truth.

These limitations mainly arise when structural cues are sparse, perspective information is ambiguous, or multiple material layers overlap within the same image region. Future work may incorporate layered depth estimation, material-aware priors, and domain-specific decorative knowledge to improve restoration reliability in complex heritage scenarios.

In a systematic review of the 220 test images, we identified 12 cases (5.4%) where the model generated plausible but historically inaccurate patterns (e.g., replacing geometric motifs with floral ones). All such cases occurred when the damaged area exceeded 40% of the image and when fewer than three structural cues (edges/depth) remained. This suggests a clear direction for future work: incorporating a knowledge base of period-appropriate decorative motifs to constrain generation when structural priors are weak.

4.9 section Summary

This section has evaluated the SACGF framework through comprehensive qualitative and quantitative experiments. The results demonstrate that the joint use of Canny edge maps, depth maps, and SAM semantic masks substantially improves structural consistency, texture authenticity, and style coherence in lacquerware surface restoration. The ablation study further demonstrates the complementary contributions of individual condition modules to overall restoration performance. Quantitative metrics on the 2,194-image dataset show that SACGF consistently outperforms CycleGAN, SD+LoRA, and their progressively constrained variants across PSNR, SSIM, LPIPS, CIEDE2000, EdgeF, and the composite DLFS. The analysis of failure cases identifies limitations in handling overlapping fine decorations and severely degraded regions, pointing to concrete directions for future improvement.

5. Discussion

This section presents a systematic discussion of the experimental results obtained by the proposed SACGF framework on the task of digital lacquerware surface restoration. It first analyzes the contribution mechanisms of each condition module to restoration performance, then situates this study within the broader context of cultural heritage digitization and compares it with existing methods, followed by a candid examination of current limitations, and concludes with an outlook on future research directions.

5.1 Effectiveness Analysis of Condition Control Modules

The ablation study presented in section 4 (Table 5 and Figure 12) clearly demonstrates that each introduced condition control module yields substantial and quantifiable improvements in restoration quality.

5.1.1 Role of the Edge Condition (Canny)

When only the LoRA style adapter is used, the model can generate images with the basic color

palette and textural appearance of lacquerware; however, the geometric orientation and continuity of cracks frequently exhibit deviations: the inpainted cracks may drift or bifurcate, failing to connect naturally with the intact cracks in the undamaged regions of the input. After incorporating our trained Canny edge ControlNet, PSNR improves by 1.39 dB and SSIM by 0.037 (Figure 11(c)). This validates that a ControlNet trained specifically on lacquer structural priors significantly outperforms generic conditioning strategies. This indicates that the edge condition provides the model with pixel-level spatial anchors, compelling the generation process to adhere strictly to the structural priors of the input image at crack positions and contour boundaries. For restoration tasks, preserving the topological consistency of crack networks is more critical than pursuing textural aesthetics alone—this is the digital manifestation of the conservation principle of "restoring the old as it was."

5.1.2 Role of the Depth Condition

With the further addition of the depth condition on top of Canny, Edge Fidelity (EdgeF) increases from 0.741 to 0.789. The value of the depth map is manifested in two respects. First, lacquerware surfaces possess complex stratified structures (e.g., the alternating black and red layers of tixi carved lacquer, the high and low relief of carved lacquer (Figure 11(d))); the depth condition enables the model to distinguish between superficial cracking and deep penetration, thereby generating shadows and a sense of layering that conform more closely to physical logic. Second, when the restoration region spans lacquer layers at different depths (e.g., from the lacquer ground to gold powder decoration), the depth map helps the model avoid flattening errors and preserve foreground-background occlusion relationships. Experiments reveal that the depth condition yields particularly significant improvements for lacquerware categories with strong three-dimensionality, such as tixi and carved lacquer.

5.1.3 Fencing Effect of Semantic Masks (SAM)

Comparing the complete framework with the "Canny + Depth" configuration, the most substantial improvements occur in LPIPS (decreasing from 0.087 to 0.062) and CIEDE2000 (decreasing from 5.91 to 4.18). These two metrics respectively measure perceptual similarity and color fidelity, and their marked improvement is attributable to the regional semantic constraints provided by SAM. Without SAM, the model occasionally allows the red of the background lacquer ground to bleed into gold-traced areas, or erroneously diffuses the luster of gold powder into crack interiors. By generating precise binary masks, SAM constructs impermeable boundaries between different semantic regions, ensuring that restoration operations are strictly confined to flaking or cracked areas without altering intact decorative patterns (Figure 11(e)). This semantic fencing effect is particularly critical for complex lacquerware incorporating multiple materials (e.g., gold, mother-of-pearl, eggshell).

5.2 Comparison with Existing Methods and a Re-examination of "Cultural Fidelity"

5.2.1 Beyond Pixel-Level and Perceptual Restoration

Traditional image restoration methods—such as partial differential equation (PDE)-based inpainting and PatchMatch-based texture synthesis—and early deep learning approaches like CycleGAN primarily focus on the plausibility of pixel values or the continuity of local textures, while neglecting the materiality and craft logic unique to lacquer art. The experimental results in this study show that CycleGAN achieves a PSNR of only 21.34 dB on lacquerware restoration and frequently produces "hallucinated decorations"—for instance, rendering ice-crack patterns as irregular grid-like motifs. This occurs because its generator attempts to minimize the adversarial loss and, in the absence of structural constraints, tends to "invent" textures that appear plausible but

are culturally incorrect.

By contrast, the proposed framework explicitly injects multimodal structural conditions (edge, depth, semantic), elevating the restoration process from pixel fitting to structural reshaping. The model learns not only what color a crack should be, but also along which direction it should extend and how it varies in depth. This shift constitutes the critical transition from “visual resemblance” to “cultural fidelity.”

5.2.2 Data Efficiency and Domain Adaptation

The Chu-style lacquerware dataset (CSL) constructed by Bi et al. and the HITL-AIGC system proposed by Liang et al. both confirm the decisive role of domain-specific datasets in fine-tuning generative models[41,43]. The present study advances this line of work: it not only uses 2,194 intact images for LoRA fine-tuning, but also proactively synthesizes damaged–intact image pairs, enabling the model to learn the restoration mapping under supervised conditions. This strategy of “self-supervised synthesis + authentic style transfer” effectively alleviates the extreme scarcity of paired data in the cultural heritage domain. Experiments demonstrate that, with only slightly over a thousand images and several hours of GPU training, a model capable of practical crack completion and flake restoration can be obtained.

5.2.3 A Quantifiable Pathway toward “Cultural Fidelity”

As noted in Section 1, existing AI-driven cultural heritage restoration research lacks a quantifiable evaluation system for “cultural fidelity.” The Digital Lacquer Fidelity Score (DLFS) and its constituent metrics (EdgeF, GLCM statistics, CIEDE2000, etc.) proposed in this paper provide, for the first time, an operational technical pathway for this abstract concept. Through DLFS, this study can objectively compare the restoration quality of different methods and diagnose whether specific failure cases stem from structural distortion, texture deviation, or color error. Although this evaluation protocol is customized for lacquer art, its underlying idea—decomposing cultural characteristics into measurable physical statistics—can be extended to the digital preservation of other intangible cultural heritage, such as ceramic crackle patterns and embroidery stitch techniques. Figure 16 presents a DLFS radar comparison of different methods, illustrating how the proposed evaluation system distinguishes structural, textural, and color fidelity in a single composite view.

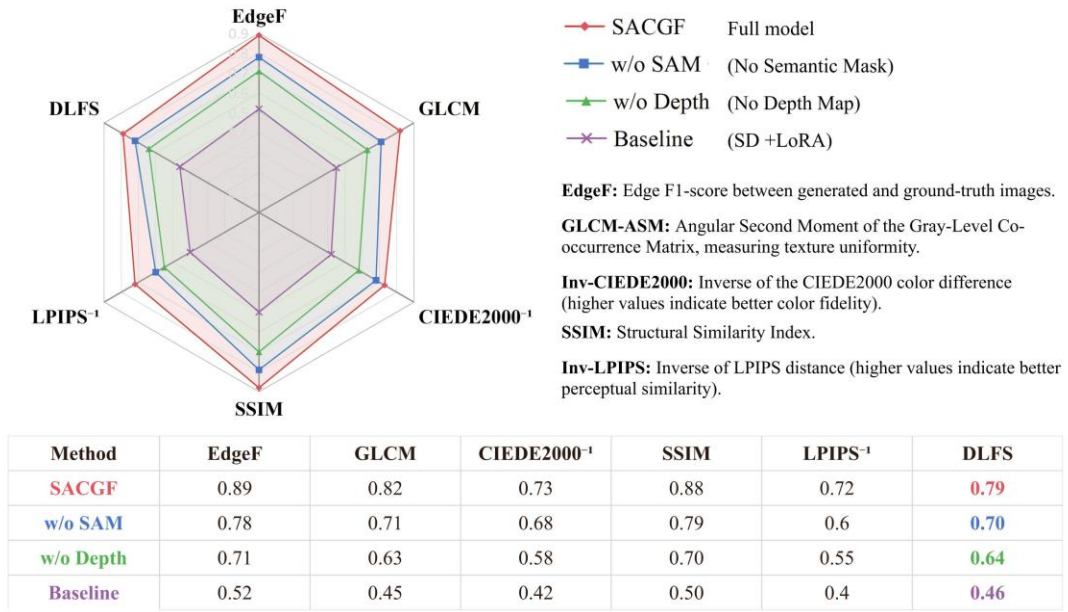


Fig16. DLFS Radar Comparison

5.4 Implications for Digital Preservation of Cultural Heritage

Beyond presenting a technical solution, this study offers three conceptual insights for the digital preservation of cultural heritage.

5.4.1 From “Recording” to “Reconstruction”: A Paradigm Shift

The core of traditional digital preservation is high-fidelity recording—preserving the current state of artifacts as non-destructively as possible through 3D scanning, multispectral photography, and similar means. This study demonstrates an alternative possibility: active reconstruction based on generative models. It not only records the morphology of cracks, but also “repairs” them in digital space and quantitatively evaluates the cultural fidelity of the restoration results. This enables heritage conservators to conduct reversible virtual restoration simulations before physical intervention, assessing the long-term effects of different treatment strategies. Such a “simulate first, implement later” workflow can substantially reduce the risks associated with actual restoration.

5.4.2 A Pathway for Integrating AI and Artisan Knowledge

The HITL-AIGC system has already demonstrated the effectiveness of human–AI collaboration[43]. This study further shows that structural conditions (edge, depth) serve as an ideal bridge between artisan knowledge and AI models. Artisans do not need to write code; they only need to provide a hand-drawn sketch of crack orientation or a reference photograph of an artifact, and the proposed framework can automatically extract edge and depth conditions to generate a restoration plan aligned with the artisan’s intent. This “sketch-as-code” mode of interaction holds promise as a new paradigm for the digital transmission of intangible cultural heritage.

5.4.3 The Importance of Explainable Evaluation

The cultural heritage field harbors an inherent wariness toward “black-box” models[44]. The DLFS evaluation system established in this study renders the model’s decision-making process transparent by decomposing cultural fidelity into interpretable physical statistics (e.g., GLCM contrast, edge retention). When restoration quality falls short, the component metrics of DLFS can pinpoint the source of the problem—whether the texture is overly smooth (elevated LPIPS) or

whether there is a color deviation (excessive CIEDE2000). Such explainable evaluation is a prerequisite for AI to earn the trust of cultural heritage conservators.

5.6 Section Summary

This section has provided an in-depth discussion of the effectiveness mechanisms of each condition module based on the experimental results of the SACGF framework, and has compared the proposed approach with existing cultural heritage digitization methods, clarifying the paradigm shift from “visual similarity” to “cultural fidelity.” It has candidly analyzed the limitations of the method in handling complex overlapping decorations, severely degraded regions, and computational resource demands, and has proposed corresponding solutions. Finally, from the perspectives of workflow, human–AI collaboration, and explainability, this section has summarized the implications of this research for cultural heritage preservation, and has outlined future directions in the integration of physical priors, three-dimensional restoration, zero-shot learning, and lightweight deployment. These discussions are intended to serve as a valuable reference for subsequent research in this field.

6. Conclusion and Future Work

This study focuses on the digital preservation of Chinese lacquer art as an intangible cultural heritage and proposes and validates a structure-aware restoration framework for lacquer surfaces. The aesthetic essence of lacquer art is rooted in its complex layered structures, stochastic crack networks, and warm luster that shifts with light and shadow—material properties that have been systematically overlooked in existing digital reconstruction methods, resulting in a fundamental disconnect between “visual similarity” and “cultural fidelity.” The central argument of this paper is that high-fidelity cultural reconstruction cannot rely on a single “black-box” generative model; rather, structure, texture, and semantics must be jointly injected as explicit priors into a controllable restoration process.

In pursuit of this research objective, this paper has constructed a corresponding technical framework, an evaluation system, and experimentally validated its feasibility in the task of digital lacquerware restoration.

At the conceptual level, this paper formalizes digital lacquer restoration as a multi-objective optimization problem that simultaneously addresses structural, textural, and stylistic constraints. Distinct from generic image inpainting or style transfer tasks, it proposes that restoration quality should be jointly measured along three orthogonal dimensions—structural consistency, texture fidelity, and stylistic coherence—and abstracts this into a multi-objective optimization problem. This formalization provides a clear constraint logic for subsequent technical choices: any effective restoration method must respect geometric boundaries, material logic, and historical aesthetics; none of the three can be absent.

At the technical level, this paper designs the Structure-Aware Collaborative Generative Framework (SACGF). Built upon Stable Diffusion as the generative backbone, the framework achieves controllable restoration through three collaborative modules: LoRA injects domain-specific stylistic features into over a thousand lacquerware images with minimal parameter overhead; A purpose-trained, multi-condition ControlNet extracts edge, depth, and color priors from the input image to impose pixel-level structural regularization on the denoising process; and SAM provides semantic masks that construct impermeable boundaries between different material regions.

The collaborative logic of the three is as follows: LoRA addresses “where the texture comes from” by learning the stylistic characteristics of lacquer surfaces; ControlNet supplies structural and spatial constraints; and SAM reduces cross-region texture diffusion and color bleeding through regional masking. This division of labor enables the framework to output reconstruction results that possess both craft logic and visual plausibility even under data-scarce conditions.

At the evaluation level, this paper establishes a multi-dimensional quantitative protocol oriented toward “cultural fidelity.” Traditional metrics such as PSNR and SSIM cannot reflect the crack sharpness, layer thickness, and gloss contrast unique to lacquer art. The Digital Lacquer Fidelity Score (DLFS) proposed in this paper deconstructs the abstract notion of “cultural fidelity” into three computable dimensions—texture conformity, color fidelity, and structural similarity—operationalized through metrics such as LPIPS, GLCM statistics, CIEDE2000 color difference, and edge retention. This evaluation protocol provides a unified quantitative standard for the experiments in this paper and also offers a reusable evaluation paradigm for subsequent research on digital restoration of cultural heritage.

At the empirical level, this paper conducted systematic experiments on a self-constructed dataset of 2,194 lacquer art images. Quantitative results show that the complete framework outperforms the compared methods on metrics including PSNR, SSIM, LPIPS, CIEDE2000, and EdgeF. Ablation studies confirm that each condition module yields consistent performance gains, with Canny edge control contributing the most to structural fidelity and SAM semantic masks contributing most prominently to the precise constraint of color and texture boundaries. Visual comparisons demonstrate that the framework can successfully complete the ice-crack network of carved red lacquer, restore the gold powder luster of black lacquer with gold tracing, reconstruct the fine random texture of ox-hair cracks on plain lacquer, and reproduce the sharp alternating black-and-red cross-sections in layered structures such as *taxi*. The analysis of failure cases further reveals the current method’s limitations in dealing with complex overlapping decorations and severely degraded regions, providing clear directions for future improvement.

6.1 Limitations

Although SACGF performs well on the majority of test cases, this paper soberly recognizes that the current method still has several fundamental limitations. The damage patterns of real artifacts are more complex than those in synthetic data, and domain adaptation using real artifact images is needed in the future.

6.1.1 Difficulty in Restoring Complex Overlapping Decorations

As described in Section 4.7, when cracks highly overlap with fine decorations such as gold tracing and mother-of-pearl inlays, the model occasionally blurs the decorative boundaries or even incorrectly transfers gold powder textures into the crack interior. An in-depth analysis reveals the cause: while the current SAM semantic masks can segment “gold powder regions” and “crack regions,” when the two overlap in a two-dimensional image (the gold powder is actually physically above or below the crack), the model lacks an understanding of the three-dimensional stacking relationship. In other words, the depth condition used in this paper is relative depth (distance from the camera), rather than fine-grained “inter-layer ordering” (e.g., the gold foil layer lies above the lacquer layer). Future work may introduce layered depth estimation or optical flow fields to explicitly model the up-down relationships among different material layers.

6.1.2 “Hallucination Risk” in Severely Degraded Regions

When more than 40% of the lacquer surface in the input image is severely flaked and very few

structural cues remain (e.g., only a small corner survives), the model may produce completions that are plausible but historically inaccurate—i.e., “hallucinations.” For example, in Figure 4.5 (right), the model filled a completely flaked region with abstract cloud motifs, whereas the original decoration of the artifact was actually geometric patterns. Such hallucinations are unacceptable in cultural heritage restoration because the goal of digital restoration is to recover the authentic original appearance, not “creative completion.” Possible remedies include: (1) constructing a lacquer decorative pattern knowledge base (e.g., pattern regularities for specific dynasties and vessel types) to constrain generation through a knowledge graph; (2) introducing “uncertainty estimation,” so that when the model has low confidence in the completed content, it outputs a mask indicating that manual intervention is needed.

6.1.3 Generalization to Real Artifact Damage

The damaged dataset in this paper was generated through procedural synthesis, whereas the deterioration patterns of real artifacts (natural aging, microbial corrosion, human-induced damage) are far more complex. Although the synthetic data simulate common crack morphologies (snake-belly, ox-hair, ice cracks) and flaking shapes, they may fail to cover certain rare or region-specific degradation modes. Therefore, for application to actual artifact restoration, the model first requires domain-adaptive fine-tuning on a small number of real samples to achieve ideal performance.

6.2 Future Work

Based on the above discussion, this paper proposes several research directions worthy of further exploration.

6.2.1 Incorporating Physical Simulation Priors

The current method relies primarily on data-driven texture learning and insufficiently models physical processes such as lacquer film drying shrinkage and multi-layer curing. In the future, physical simulation methods such as finite element analysis could be embedded into the generation pipeline: first predicting crack growth paths through physical models, and then rendering material details via the diffusion model. Such a hybrid “physics + learning” framework holds the potential to fundamentally resolve the hallucination problem.

6.2.2 Restoration of Three-Dimensional Lacquer Surfaces

This study only addresses two-dimensional images. However, the beauty of lacquerware is inherently three-dimensional—the three-dimensional tool marks of carved lacquer, the inlay depth of mother-of-pearl, and the viewing-angle dependence of gloss. The next step will be to extend SACGF to three-dimensional meshes or Neural Radiance Fields (NeRF), achieving a leap from “2D image restoration” to “3D surface reconstruction.”

6.2.3 Cross-Category Zero-Shot Restoration

The current training set covers eight major lacquerware categories, but Chinese lacquer art includes many more rare types (e.g., baibaoxian multi-treasure inlay, kuancai polychrome engraving). Can a unified restoration model be trained that is capable of zero-shot restoration on unseen categories? This requires the design of more general structural representations (e.g., abstracting cracks as curve graphs, mapping materials as vector fields) rather than relying on category-specific texture memory.

6.2.4 Lightweight Deployment and Mobile Applications

This paper plans to compress the inference model to under 100 MB through model quantization (INT8) and knowledge distillation, enabling it to run on smartphones. This would allow field archaeologists to perform virtual reassembly and restoration of excavated lacquerware fragments on site, greatly improving work efficiency.

6.3 Research Limitations and Future Directions

The method presented in this paper is subject to three main constraints. First, for inputs with over 40% area loss and few structural cues, the model may produce “hallucinated” decorations that are semantically plausible but historically inauthentic. Second, when cracks highly overlap with fine decorations such as gold tracing and mother-of-pearl, two-dimensional semantic masks struggle to distinguish the stacking order, leading to boundary blurring. Third, the current framework relies on high-performance GPUs, limiting direct deployment in field archaeology and grassroots heritage conservation units. Future work will proceed along three paths: first, introducing physical priors such as finite element analysis to explicitly model crack growth paths and suppress hallucinations at their root; second, extending the framework from two-dimensional images to three-dimensional meshes or Neural Radiance Fields for full geometric reconstruction of lacquer surfaces; and third, developing a lightweight mobile version through model quantization and knowledge distillation, making digital restoration tools truly accessible.

In summary, this study demonstrates that, under relatively limited data conditions, combining structural constraints, style adaptation, and semantic guidance can effectively enhance the stability and fidelity of digital lacquer surface restoration. What the digital transmission of lacquer art requires is not more complex generative models, but deep respect for material logic and quantifiable criteria for cultural fidelity. The SACGF framework and its evaluation system provide an operable, reproducible, and improvable starting point for this direction. In the future, with the further integration of generative models, physical simulation, and cultural knowledge bases, artificial intelligence is poised to play a more significant role in the digital preservation and assisted restoration of cultural heritage.

Author Contributions Statement : Implemented the concepts, methods, and software applications based on specific implementation suggestions, and authored the paper. The authors declare that they have no known competing financial interests or personal relationships.

Data Availability Statement : The lacquer art image dataset constructed and used in this study is not publicly available due to the inclusion of images sourced from copyrighted publications, museum collections with restricted access policies, and materials provided by intangible cultural heritage inheritors under specific usage agreements. The dataset description, including the complete list of data sources, hierarchical classification schema, and preprocessing pipeline, is fully documented in Section 3.6.1. Researchers interested in accessing the dataset for non-commercial academic purposes may contact the corresponding author with a formal request, subject to the approval of the original data providers.

Conflict of interest: the authors declare no conflict of interest.

Ethics Review: This study has been reviewed and approved by the Ethics Committee of Jiangxi Vocational and Technical College of Communications for conducting user questionnaire surveys. Review Number: IRBJVT20260321. All participants in this survey provided informed and voluntary consent prior to participation.

Funding Source: This study was fully funded by the Jiangxi Provincial Vocational Education Teaching Reform Research Project—General Category—“Research and Practice of the ‘Virtual-Physical-Industrial’ Teaching Model for the ‘Labor Education’ Course in Vocational Colleges Empowered by Intangible Cultural Heritage” (Project No.: JXJG-24-53-11).

References

- [1] DongYu, R., Bin Harun, A., & YiXuan, L. (2024). The Light of Chinese Lacquerware-Dynastic Contribution and Exploration of Cultural Connotation. *Online Journal of Art & Design*, 12(4).
- [2] Rao, D., Harun, A., & Li, Y. (2025). Explore Chinese lacquer art culture—eggshell inlay technique/Rao DongYu, Prof. Madya Dr. Azahar Harun and Li YiXuan. *Voice of Academia (VOA)*, 21(1), 124-139.
- [3] Zhang, W., & Mohamed, F. N. (2025). Student Group’s Perception of Cultural Symbols in Lacquerware Enhances Cultural Identity. *Journal of Literature Advances*, 2(1), 15-18.
- [4] Suo, C., & Noknoi, K. (2025). The Modern Conversion Type of Cultural Identity and Literacy Role in the Context of Traditional Craft Revitalization: The Case of Lacquerware of the Yi Ethnic Group in Liangshan, China. *International Journal of Education and Literacy Studies*, 13(1), 138-146.
- [5] Zhang, A., Gong, Y., Chen, Q., Jin, X., Mu, Y., & Lu, Y. (2025). Driving innovation and sustainable development in cultural heritage education through digital transformation: The role of interactive technologies. *Sustainability*, 17(1), 314.
- [6] Wang, X., & Chen, B. (2025). Artificial intelligence for cultural heritage: digital image processing-based techniques and research challenges. *International Journal of Information and Communication Technology*, 26(13), 37-60.
- [7] Zhao, L., Chen, Y., Du, G., & Wu, X. (2025). Semantic and Sketch-Guided Diffusion Model for Fine-Grained Restoration of Damaged Ancient Paintings. *Electronics*, 14(21), 4187.
- [8] DePolo, G., Walton, M., Keune, K., & Shull, K. R. (2021). After the paint has dried: a review of testing techniques for studying the mechanical properties of artists’ paint. *Heritage Science*, 9(1), 1-24.
- [9] Guo, J., & Ahn, B. (2023). Tacit knowledge sharing for enhancing the sustainability of intangible cultural heritage (ICH) crafts: a perspective from artisans and academics under craft–design collaboration. *Sustainability*, 15(20), 14955.
- [10] HU, J., GENG, G., XIONG, M., LI, S., & ZHANG, Y. (2025). A Style Transfer Method for Chinese Landscape Painting Based on Detail Feature Extraction and Fusion. *Advanced Engineering Sciences*, 57(1), 98-106.
- [11] Liang, J., Wei, J., & Jiang, Z. (2020). Generative adversarial networks GAN overview. *Journal of Frontiers of Computer Science and Technology*, 14(1), 1-17.
- [12] Li, J., Wu, L., Xu, D., & Yao, S. (2023). Soft multimodal style transfer via optimal transport. *Knowledge-Based Systems*, 271, 110542.
- [13] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1833-1844).
- [14] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36, 10088-10115.
- [15] Guo, H., Dai, T., Bai, Y., Chen, B., Ren, X., Zhu, Z., & Xia, S. T. (2024). Parameter

efficient adaptation for image restoration with heterogeneous mixture-of-experts. *Advances in Neural Information Processing Systems*, 37, 13522-13547.

[16] Guo, J., Li, J., Fu, H., Gong, M., Zhang, K., & Tao, D. (2022). Alleviating semantics distortion in unsupervised low-level image-to-image translation via structure consistency constraint. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18249-18259).

[17] Jiang, J., Han, Q., Zhang, C., & Zheng, Z. (2026). High-Fidelity Mural Restoration via a Unified Hybrid Mask-Aware Transformer. *arXiv preprint arXiv:2604.03984*.

[18] Zhang, J. (2019). New functionalism: rejuvenating historical and cultural heritage through urban revival. *International Journal of Anthropology and Ethnology*, 3(1), 8.

[19] Sameera M Al-Ghamdi, Nadia Yusuf, Digital heritage and fashion: preserving cultural identity through AI and virtual archives, *Digital Scholarship in the Humanities*, 2026;, fqag008, <https://doi.org/10.1093/llc/fqag008>

[20] Xu, J., Yan, L., Zhang, R. et al. A review of the development and application of generative technology in digital museums. *npj Herit. Sci.* 13, 589 (2025). <https://doi.org/10.1038/s40494-025-02164-1>

[21] Yu, Y. (2010). *Gudai Qiqi (Ancient Lacquerware)*. Jilin Wenshi Press.

[22] Shen, F., & Li, D. (1984). *Qiqi Gongyi Jifa Jiye (Techniques of Lacquerware Craft)*. Bei

[23] Wang, S. (1998). *Xiushi lu jieshuo: Zhongguo chuantong qi gongyi yanjiu* (Rev. ed.) [Annotated and explained of the Xiushi Lu: A study on Chinese traditional lacquer craftsmanship]. Cultural Relics Publishing House.

[24] Zhu, C. (2016). *Chinese lacquer art and design*. China Architecture & Building Press.

[25] Tie, Y. (Ed.). (2005). *Zhongguo gudai qiqi [Ancient Chinese lacquerware]*. Hualing Publishing House.

[26] Zhao, Y. (2016). *Chushi qiqi [Chu-style lacquerware]*. Hubei People's Press.

[27] Chen, C. (2015). *Qihua chuanguzuo jifa [Lacquer painting creation techniques]*. Dalian University of Technology Press.

[28] Zhu, C. (2016). *Zhongguo qi yi yu sheji [Chinese lacquer art and design]*. China Architecture & Building Press.

[29] Tai, G. (2021). *Qiqi [Lacquerware]*. Chongqing Publishing House.

[30] Feng, W., Zhang, Q., Tian, F., Wang, X., Chai, B., Sun, J., & Su, B. (2021). Mianxiang yufangxing baohu de wenwu benti zhineng yuanwei jiance xitong [An intelligent in-situ visual inspection system for preventive conservation of cultural heritages]. *Scientia Sinica Informationis*, 51(12), 2102.

[31] Abdel-Aall, O., Abdel-Ghani, M., & Abd El-Monem, H. (2025). CONSERVING OF PERSIAN LACQUER PAINTING, MUSEUM OF ISLAMIC ART, CAIRO. *Egyptian Journal of Archaeological and Restoration Studies*, 15(2), 181-188.

[32] Wang, B., Wang, J., Bao, Q., & Liang, N. (2024). RESEARCH ON AN ADAPTIVE IMAGE ENHANCEMENT ALGORITHM FOR LACQUER PAINTING IMAGES IN LOW-LIGHT ENVIRONMENTS BASED ON THE SPARROW SEARCH ALGORITHM AND INCOMPLETE BETA FUNCTION. In *DS 136: Proceedings of the Asia Design and Innovation Conference (ADIC) 2024* (pp. 102-110).

[33] Spennemann, D. (2025). Wither Heritage? How Does a Generative Artificial Intelligence Model Envision Humanity's Cultural Heritage in 30 Years' Time?. Available at SSRN 5201057.

- [34] Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2414-2423).
- [35] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2223–2232.
- [36] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.
- [37] Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3836–3847.
- [38] Dunhuang mural restoration using GAN Ren, H., Sun, K., Zhao, F., & Zhu, X. (2024). Dunhuang murals image restoration method based on generative adversarial network. *npj Heritage Science*, 12(39).
- [39] Wang, Q., Jiang, C., Lu, Z., Liu, X., Jiang, K., & Liu, F. (2025). Sustainable conservation of embroidery cultural heritage: An approach to embroidery fabric restoration based on improved U-Net and multiscale discriminators. *Applied Sciences*, 15(19), 10397.
- [40] Rao, Y., Chen, S., Xuan, Y., Hu, B., Wang, R., & Li, M. (2026). Image generation method for Cantonese embroidery artistic style based on diffusion models. *npj Heritage Science*, 14, 79. <https://doi.org/10.1038/s40494-026-02342-9>
- [41] Bi, H., Chen, Y., Chen, C., & Shu, L. (2025). Chu-Style Lacquerware Dataset: A Dataset for Digital Preservation and Inheritance of Chu-Style Lacquerware. *Sensors*, 25(17), 5558.
- [42] Wang, J. (2025). Geometric morphology optimization method based on image recognition technology in lacquerware pattern designs. *International Journal for Housing Science and Its Applications*, 46(3), 6734–6744. <https://doi.org/10.70517/ijhsa463578>
- [43] Liang, J., Li, Y., Xiong, Z., & Huang, Q. (2025). Advancing lacquerware design through human-AI collaboration with controllable diffusion models. *Scientific Reports*. <https://doi.org/10.1038/s41598-025-33119-y>
- [44] Park, S. (2025, September). Revisiting Transparency in AI-based Creation: Unpacking ‘Cultural Black Boxes’ towards ‘Cultural Explainability’. In *2025 IEEE International Symposium on Technology and Society (ISTAS)* (pp. 1-7). IEEE.