

A Heritage-Aware Evaluation Framework for Synthetic Training Images of Kizil Grotto Murals

NA XU

Beijing University of Posts and Telecommunications

XUAN ZHOU

Beijing University of Posts and Telecommunications

LIHUI MIAO

Kizil Caves Research Institute of Xinjiang

HAIYING ZHAO

zhaohaiping@bupt.edu.cn

Beijing University of Posts and Telecommunications

KUN XU

Beijing University of Posts and Telecommunications

Article

Keywords: Kizil Grotto murals, synthetic training images, heritage-aware evaluation, generative models, quality control, digital heritage

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11134480/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Heritage-Aware Evaluation Framework for Synthetic Training Images of Kizil Grotto Murals

Na Xu^{1,2}, Xuan Zhou^{1,2}, Lihui Miao³, Haiying Zhao^{1,2*}, Kun Xu^{4*}

¹School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, 100876, China.

²Beijing Key Laboratory of Cultural Intelligence Computing, Beijing, 100876, China.

³Kizil Caves Research Institute of Xinjiang, Kuqa, Xinjiang, 842000, China.

⁴State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing, 100876, China.

*Corresponding author(s). E-mail(s): zhaohaiying@bupt.edu.cn;
xukun@bupt.edu.cn;

Abstract

Generative models offer a practical means of expanding limited cultural heritage image datasets, but unfiltered synthetic images may introduce semantic drift, compositional distortion, color deviation and line artefacts into training corpora. We developed a heritage-aware computational framework to evaluate synthetic Kizil Grotto mural images and support their screening for training datasets. The framework integrates ten indicators across semantics, composition, color and line, using criterion weights derived through the analytic hierarchy process. Using 24 reference images depicting non-human motifs, selected from 94 manually restored images derived from 1,479 field photographs of Cave 38, we evaluated 10,119 images generated by Stable Diffusion, Midjourney, DALL·E 3 and Kandinsky. Semantics received the highest weight (0.3982), followed by composition (0.2804), color (0.2019) and line (0.1195). Under the tested conditions, Midjourney achieved the highest overall score (0.713), although performance varied across indicators. The framework supports comparative assessment and heritage-oriented curation of candidate synthetic training images.

Keywords

Kizil Grotto murals; synthetic training images; heritage-aware evaluation; generative models; quality control; digital heritage

1. Introduction

Kizil Grotto murals constitute an important corpus of Kucha Buddhist art along the Silk Road and preserve substantial evidence of religious narrative, artistic exchange and mural-painting practice¹. However, deterioration, conservation restrictions and limited physical access constrain the quantity and diversity of mural images available for computational research and model training². Generative models provide a practical means of expanding such limited datasets, but unfiltered synthetic images may introduce semantic errors, stylistic drift, color distortion, structural artefacts and model-specific biases. When incorporated into training datasets, these defects may be learned or amplified by downstream restoration and recognition models. Synthetic data augmentation for cultural heritage therefore requires not only large-scale image generation but also heritage-specific quality control.

Kizil Grotto murals are renowned for their distinctive artistic style. Their narrative subjects primarily include Jātaka tales and Avadāna tales (Fig. 1a). Their most characteristic compositional feature is the diamond-grid scheme, in which the vaulted surface is divided into numerous diamond-shaped compartments, each depicting a representative moment from an independent story (Fig. 1b)³. Their vivid palette is dominated by mineral pigments such as lapis lazuli blue, atacamite green and

hematite red, producing strong color contrasts and rich chromatic layers (Fig. 1c)⁴. Their lines show relatively little variation in width and are typically composed of repeated U-shaped curves, creating a firm, regular and decorative linear rhythm (Fig. 1d)⁵. Together, these narrative, compositional, color and line characteristics provide heritage-specific criteria for evaluating whether synthetic training images retain the visual identity of Kizil Grotto murals.

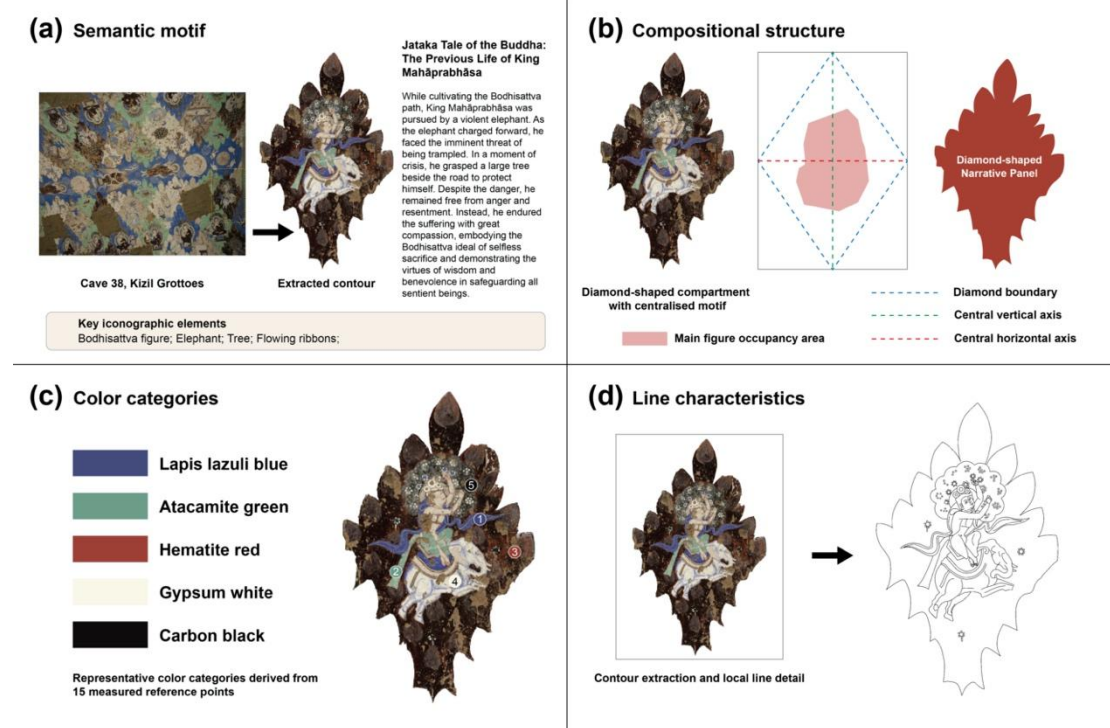


Fig. 1 | Heritage-specific visual characteristics of Cave 38 murals.

Neural style transfer, image-to-image translation and diffusion-based generation offer technical pathways for transferring, reconstructing and reinterpreting visual style. Early neural style transfer methods demonstrated that content and style representations could be separated and recombined through deep convolutional features⁶. Subsequent GAN-based and cycle-consistency-based models enabled image translation across paired or unpaired domains^{7,8}, while recent diffusion models and text-to-image systems have improved the controllability, resolution and semantic responsiveness of stylised image generation⁹. These developments make it possible to produce large numbers of visually varied images from a comparatively limited set of cultural heritage references.

In mural heritage research, style-transfer-based algorithms have gradually become an important approach for improving the accuracy and reliability of restoration models, because they can expand training samples, simulate degradation features such as cracks and fading, and provide quantitative criteria for evaluating restoration results². These methods have increasingly been applied to cultural heritage images, particularly in mural restoration and mural inpainting, color restoration and mural style transfer. In mural restoration and inpainting, generative models have been used to reconstruct missing or damaged regions by combining semantic prediction, structural guidance and texture synthesis¹⁰⁻¹². Beyond restoration, related studies have extended generative models to mural style transfer and heritage-oriented image synthesis. Representative examples include Dunhuang mural style transfer based on generative adversarial networks¹³, tomb mural style transfer using feature separation and fusion², Dunhuang art style transfer using hierarchical vision transformers and color-consistency constraints¹⁴, and Thangka mural style transfer based on progressive style-attentional networks and multi-level loss functions¹⁵. These studies show that generative models can support digital restoration, visual reconstruction and style-based visual reinterpretation of mural heritage^{16,17}.

Although these studies demonstrate the capacity of generative models to produce mural-related imagery, they primarily evaluate generation or restoration performance rather than the suitability of synthetic outputs for inclusion in heritage training datasets. Evaluation commonly focuses on model

architecture, restoration accuracy, content preservation, artefact reduction, color consistency and overall style similarity^{18,19}. Although these criteria are important for assessing technical performance, they are insufficient for screening synthetic training images, because a visually plausible output may still contain semantic errors, compositional deviations, color distortion or line irregularities that weaken the identity of the source mural tradition. In the case of Kizil Grotto murals, relevant characteristics include the semantic representation of Jātaka and Avadāna narratives, diamond-grid compositional organisation, mineral-pigment-based color relationships and relatively uniform lines formed by repeated U-shaped curves^{3-5,20-22}.

General image-quality, perceptual-similarity and image-text alignment metrics, such as the structural similarity index measure (SSIM), Fréchet inception distance (FID), learned perceptual image patch similarity (LPIPS) and CLIP-based similarity, provide useful measures of structural correspondence, distributional quality, perceptual similarity and semantic alignment²³⁻²⁶. Computational-aesthetic and multi-criteria approaches have further extended image assessment to semantic, stylistic, compositional and conservation-related attributes²⁷⁻³⁰. However, these methods do not directly determine whether generated images preserve the integrated characteristics of a specific mural tradition or whether they are suitable for inclusion in a heritage training corpus.

Recent studies have begun to develop domain-specific approaches for evaluating AI-generated cultural images, including perceptual and structure-aware assessment of generated Thangka images and culturally grounded rubrics for AI-generated cultural artefacts^{31,32}. However, these studies primarily address generated-image quality or cultural appropriateness rather than the curation of synthetic training data. The use of heritage-specific evaluation as a quality-control mechanism for screening candidate training images therefore remains insufficiently investigated. To our knowledge, no established framework has integrated the semantic, compositional, chromatic and linear characteristics of Kizil Grotto murals into an interpretable procedure for determining which synthetic images should be retained, referred for expert review or excluded from a training corpus.

We develop a heritage-aware evaluation framework for assessing and screening synthetic Kizil Grotto mural images intended for inclusion in a training-image corpus. The framework operationalises mural-specific cultural and visual characteristics through four measurable dimensions—semantics, composition, color and line—and ten computational indicators. We apply the framework to 10,119 candidate outputs generated by Stable Diffusion, Midjourney, DALL·E 3 and Kandinsky, using 24 reference images depicting non-human motifs selected from a pool of 94 manually restored images derived from 1,479 field-captured images of Cave 38. Computational measurements were integrated with Analytic Hierarchy Process-derived criterion weights to evaluate the relative suitability of the generated images as candidate training assets. The comparison is not intended to establish a universal ranking of generative systems, but to examine their relative capacity to produce images that satisfy the heritage-specific requirements of the proposed quality-control procedure.

The contributions of this study are threefold. First, we propose a heritage-aware evaluation framework that translates the cultural and visual characteristics of Kizil Grotto murals into measurable semantic, compositional, color and line indicators. Second, we establish a curated image pipeline comprising 1,479 field-captured mural images, 94 manually restored images and 24 reference images depicting non-human motifs, and use the final reference subset to generate and assess a large corpus of candidate synthetic training images. Third, we integrate computational indicators and AHP-derived criterion weighting into a structured procedure for comparing, screening and curating synthetic mural images before their potential inclusion in training datasets. In this way, the study shifts the evaluation of mural image generation from generic visual-quality comparison towards heritage-specific quality control for synthetic training data.

2. Methods

2.1 Feature dimensions for heritage-aware evaluation

The proposed evaluation framework was constructed by operationalising the visual and semantic characteristics of Kizil Grotto murals into measurable feature dimensions. Rather than treating the evaluation of synthetic mural images solely as a general image-quality problem, we defined the criteria according to the heritage-specific characteristics of Kizil Grotto mural painting and their relevance to the assessment of candidate training images. Four feature dimensions were selected: semantics, composition, color and line. These dimensions correspond to the main cultural,

structural and stylistic properties used to assess the extent to which a synthetic image retains recognisable characteristics of Kizil Grotto mural art.

The semantic dimension was used to evaluate whether a generated image preserved both the target motif and the cultural meaning associated with Kizil Grotto murals. This dimension included content semantic consistency, which refers to the correspondence between the generated image and both the target motif and the required number of motif instances specified in the prompt, and cultural semantic consistency, which refers to the preservation of iconographic and cultural attributes associated with Kizil Grotto murals. The compositional dimension was used to assess the texture characteristics and spatial organisation of the generated image. It comprised texture characteristics, visual balance and centroid position, which respectively reflected texture organisation, the distribution of visual information and the placement of the principal motif. These indicators were selected because Kizil Grotto mural imagery commonly relies on ordered pictorial structures and balanced arrangements to organise visual content. The color dimension was used to evaluate correspondence between the generated image and the color characteristics represented in the reference murals. Five reference color categories were defined: lapis lazuli blue, atacamite green, hematite red, gypsum white and carbon black. These categories provided the basis for calculating color gamut distribution, color contrast and dominant hue consistency. The line dimension was used to assess whether the generated image retained the contour-based modelling and linear regularity associated with Kizil Grotto mural painting. It comprised line smoothness and line-width consistency, which measured contour continuity and variation in line thickness, respectively. These properties are relevant to motif clarity and the visual regularity of mural forms.

The four dimensions provided the conceptual basis for the subsequent computational indicators. The framework was designed to combine cultural-semantic correspondence with measurable visual characteristics while remaining interpretable and comparable across outputs from different generative models. This design enabled the quantitative assessment of candidate synthetic training images before their potential inclusion in a Kizil Grotto mural training corpus.

2.2 Mural image dataset

To address the limited quantity and diversity of Kizil Grotto mural images available for subsequent restoration-model training, a reference dataset was constructed from field-captured images of Cave 38 of the Kizil Grottoes in Baicheng County, Xinjiang, China. A representative subset of these images was subsequently selected and restored for generating and evaluating candidate synthetic training images intended to expand the training corpus. Cave 38 was selected because it preserves relatively complete mural remains and contains rich iconographic content representative of Kizil Grotto mural art. Cave 38, also known as the “Music Cave” or “Music and Dance Cave”²², has been dated by radiocarbon analysis to approximately AD 310 ± 80²². Its murals primarily depict Jātaka and Avadāna narratives and exhibit distinctive compositional, color and line characteristics. The mural images from Cave 38 therefore provided a suitable visual corpus for developing and evaluating the proposed heritage-aware framework.

A total of 1,479 field-captured mural images were collected from Cave 38. These images covered multiple visual categories, including Buddhist figures, human figures, animals, architectural elements, plant motifs, decorative patterns and geometric forms. From this image pool, 94 representative mural images were selected according to image completeness, motif representativeness, color richness, line clarity and compositional features. These selected images were manually restored using digital techniques by three doctoral researchers with academic backgrounds in art and design and interdisciplinary expertise in artificial intelligence. The restoration process aimed to reduce the visual influence of flaking, cracks, fading, local losses and background interference. The restored images were subsequently reviewed by domain experts to assess whether the major semantic, compositional, color and line characteristics of the source murals had been retained.

The 94 manually restored images were not treated as definitive reconstructions of the murals’ original appearance. Instead, they formed a restored image pool from which 24 images depicting non-human motifs were selected as reference images for synthetic image generation and subsequent evaluation. This design provided comparatively complete visual materials while avoiding the assumption that the restored images represented definitive historical ground truth. The study dataset therefore comprised four components: 1,479 field-captured mural images, 94 manually restored mural images, 24 reference images depicting non-human motifs and 10,119 candidate synthetic

training images generated in the subsequent experiments, as shown in Table 1. Together, these components provided the data basis for developing and applying the semantic, compositional, color and line indicators in the proposed heritage-aware evaluation framework.

Table 1 | Composition of the Kizil Grotto mural image dataset

Dataset component	Number	Function
Field-captured mural images	1,479	Initial image pool from Cave 38
Manually restored mural images	94	Restored image pool for reference-image selection
Reference images depicting non-human motifs	24	Visual inputs and comparison references for synthetic image generation and evaluation
Candidate synthetic training images	10,119	Samples for quantitative evaluation

2.3 Image acquisition and reference-image preparation

The mural images were captured using a Fujifilm GFX100S medium-format digital camera equipped with standard and telephoto lenses selected according to the spatial constraints of Cave 38. Flash photography was not used because of conservation restrictions, and all images were recorded in RAW format to preserve color information, texture and fine visual details for subsequent digital restoration, reference-image preparation and indicator computation.

To improve color consistency across image batches, color reference measurements were taken using a CHNSpec DS200 spectrophotometer. Five representative color categories were defined according to the dominant mineral-based palette of Kizil Grotto murals: blue, green, red, white and black. For each color category, three visually stable regions were selected for measurement, resulting in 15 color reference points in total. Areas with severe flaking, cracks, strong shadows, surface contamination or obvious restoration traces were avoided during sampling. The measured values were used as color reference anchors to guide subsequent color correction and palette validation, rather than as a complete pigment-material analysis. The same preprocessing principles were applied across the image set to reduce uncontrolled color shifts caused by non-flash lighting conditions.

A subset of generation reference images was selected from the 94 manually restored mural images through a two-stage procedure combining manual inspection and automated color screening. During manual inspection, images depicting non-human motifs—including animals, plants, landscape motifs and objects—were prioritised to maintain a consistent experimental scope. Images dominated by identifiable human figures, faces or limbs were excluded because figure generation may introduce anatomical and facial distortions that were not specifically addressed by the present evaluation framework. Candidate images were also required to contain a recognisable principal motif, distinguishable color regions and sufficiently clear line details for subsequent reference-guided image generation.

During the automated screening stage, each candidate image was converted to the HSV color space, and K-means clustering with $K=5$ was applied to identify five dominant color clusters. For each cluster, the pixel proportion was calculated by dividing the number of pixels assigned to that cluster by the total number of pixels in the image. A candidate image was retained only when at least three of the five color clusters each accounted for more than 10% of the image pixels. This threshold was used to retain images with sufficient chromatic diversity for subsequent generation and evaluation.

The final generation reference set comprised 24 images depicting non-human motifs and was organised according to two dimensions: motif and color. The motif categories included animals, plants, landscape motifs and objects. The color categories comprised lapis lazuli blue, atacamite green, hematite red, gypsum white and carbon black, as shown in Table 2. Each selected reference image was subsequently converted into a square format while preserving the principal motif and avoiding geometric distortion. When necessary, cropping or padding was applied to centre the

principal motif and retain relevant mural information. The file size of each image was kept below 10 MB to ensure successful upload and consistent use across Stable Diffusion, Midjourney, DALL·E 3 and Kandinsky. The 24 standardised images were used as visual inputs for reference-guided synthetic image generation and as the corresponding comparison images for reference-dependent indicator calculations and qualitative visual analysis. The complete workflow, from image acquisition and color-reference measurement to multi-model generation and heritage-aware screening, is shown in Fig. 2.

Table 2 | Reference-image selection dimensions and categories

Dimension	Categories
Motif	animals, plants, landscape motifs, objects
Color	lapis lazuli blue; atacamite green; hematite red; gypsum white; carbon black

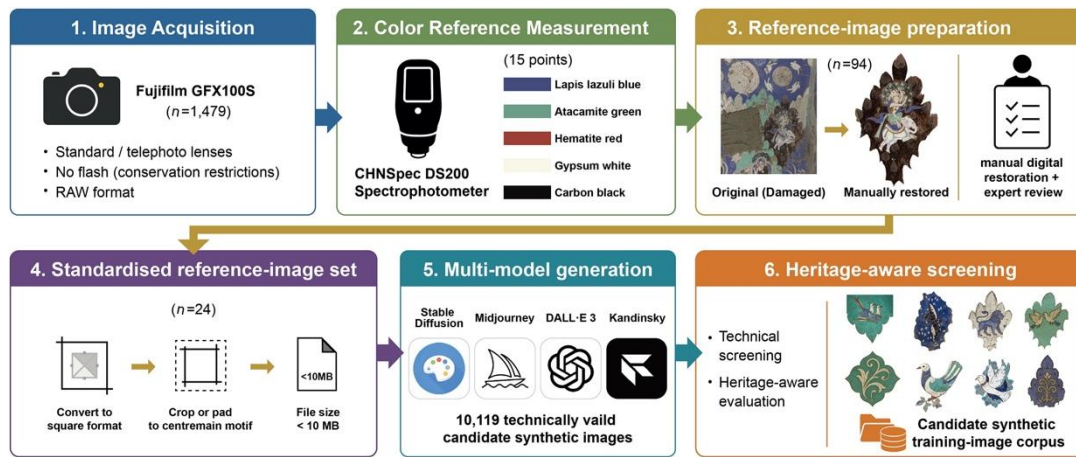


Fig. 2 | Dataset construction, preprocessing and synthetic-image generation workflow.

2.4 Generative models and prompt design

The experimental design was developed to compare different generative systems for producing candidate synthetic Kizil Grotto mural training images and incorporated model selection, prompt construction, sample generation and control variables. The experimental setup emphasised technical comparability and reproducibility, while also taking into account the distinctive characteristics of Kizil Grotto murals in terms of semantic symbols, compositional conventions, color features and line styles. This design allowed the generated results to be evaluated in relation to both artistic fidelity and cultural interpretability, and provided a basis for subsequent indicator computation and comparative evaluation of the generated outputs.

Four generative models were selected to evaluate their applicability to the production of candidate synthetic Kizil Grotto mural training images: Stable Diffusion, Midjourney, DALL·E 3 and Kandinsky. These models were selected because they differ in accessibility, control mechanisms, text-image alignment and stylised image-generation characteristics. Stable Diffusion was included as an open-source diffusion-based system with relatively flexible control and fine-tuning potential. Midjourney was included as a widely used text-to-image generation system with strong stylised image-generation capacity. DALL·E 3, developed by OpenAI, was selected because of its text-driven generation and semantic alignment capabilities. Kandinsky was included as an additional diffusion-based system with distinct stylised and abstract output characteristics. Together, these four systems provided a comparative basis for analysing how different generative approaches responded to the same Kizil Grotto mural reference images and prompt constraints.

Experimental consistency was maintained by using the same reference-image set and fixed prompt template across all four systems. Only the motif and color terms were varied across

experimental conditions. The motif and color terms followed the reference image selection categories defined in Table 2.

To ensure that the generated results maintained both stylistic consistency and artistic characteristics, we adopted a hierarchical prompt construction strategy based on a “motif + color” structure. The relationships between these prompt components and the related evaluation indicators are shown in Table 3.

The prompt template was defined as follows:

Image input: reference image R_i

Text input: “in the style of Kizil Grotto murals, preserve the basic shape and outline of the main motif in the reference image, with [N] stylized [MOTIF] motif(s) positioned within the central area of the image, forming a balanced composition, painted in a flat [COLOR], contours and details outlined in hematite-red lines, keeping the texture and aged feel of an ancient mural”

After excluding failed generations, incomplete outputs and images with severe artefacts, 10,119 valid model-generated images were retained for subsequent quantitative evaluation.

Table 3 | Prompt components and related evaluation indicators

Prompt	Indicator	Explanation
in the style of Kizil Grotto murals	Semantics: cultural semantic consistency	Specifies the style of Kizil Grotto murals, ensuring that the generated results retain the characteristic artistic and cultural features of Kizil Grotto murals and reducing style drift.
preserve the basic shape and outline of the main motif in the reference image	Composition: texture characteristics	Specifies the preservation of the basic form and outer contour of the reference motif.
with [N] stylized [MOTIF] motif(s)	Semantics: content semantic consistency	Specifies the target motif type and number.
positioned within the central area of the image	Composition: centroid position	Specifies that the motif is located at the centre of the image, corresponding to the centralised compositional convention in mural painting.
forming a balanced composition	Composition: visual balance	Encourages a balanced spatial distribution of visual information within the generated image.
painted in a flat [COLOR]	Color: color gamut distribution and dominant hue consistency	Specifies flat mineral-pigment-based coloring, avoiding modern gradients and lighting effects.
contours and details outlined in hematite red lines	Line: line smoothness and line-width consistency	Highlights the hematite red outlines commonly used in ancient murals, emphasising clear, smooth and evenly weighted lines.

keeping the texture and aged feel of an ancient mural	Texture characteristics; Cultural semantic consistency	Emphasises mural texture and a sense of historical ageing, ensuring that the generated results convey cultural heritage attributes rather than a modern illustration style.
---	---	---

2.5 Computational indicator calculation

Semantic indicators

The semantic criterion was designed to assess whether each candidate synthetic image remained consistent with both the content specified in the textual prompt and the heritage-related visual attributes represented in its corresponding reference image. In the reference-guided generation of candidate synthetic Kizil Grotto mural images, each generation condition comprised a textual prompt and a corresponding reference image. The textual prompt specified the target motif, the required number of motif instances, the color and associated visual constraints, whereas the reference image provided visual information related to the forms and cultural attributes of Kizil Grotto mural art. Accordingly, the semantic criterion comprised two computational indicators: content semantic consistency and cultural semantic consistency. Content semantic consistency evaluated the correspondence between the candidate synthetic image and the target content specified in the textual prompt, including the target motif and the required number of motif instances, whereas cultural semantic consistency evaluated its correspondence with the cultural attributes and visual identity represented in the reference image.

Content semantic consistency measured the semantic correspondence between the candidate synthetic image and the textual prompt, encompassing both global prompt–image alignment and local motif-level consistency.

Global prompt–image alignment assessed the overall semantic correspondence between the candidate synthetic image I and the complete textual prompt T . The prompt included the target motif and its specified quantity, expressed using the corresponding natural-language numeral, such as “one” or “two”, together with the remaining visual instructions. A CLIP-based multimodal model was used to encode the candidate image and the complete prompt in a shared semantic embedding space, and their cosine similarity was subsequently calculated as follows:

$$S_{\text{content}}^{\text{global}} = \cos(f_{\text{img}}(I), f_{\text{text}}(T)) \quad (1)$$

$f_{\text{img}}(I)$ and $f_{\text{text}}(T)$ denote the feature vectors produced by the CLIP image and text encoders, respectively. The cosine similarity quantified the overall semantic correspondence between the candidate synthetic image and the complete textual prompt. A value closer to 1 indicated stronger prompt–image alignment, including greater correspondence with the target motif and its specified quantity.

Local motif-level consistency assessed whether the target motif-related elements specified in the textual prompt were identifiable in the candidate synthetic image. Precision, recall and the F1 score were calculated based on the overlap between the target motif-keyword set extracted from the textual prompt T and the predicted motif-keyword set derived from the BLIP-generated caption³³:

$$P_{\text{content}} = \frac{|\mathcal{K}_T \cap \mathcal{K}_P|}{|\mathcal{K}_P|} \quad (2)$$

$$R_{\text{content}} = \frac{|\mathcal{K}_T \cap \mathcal{K}_P|}{|\mathcal{K}_T|} \quad (3)$$

$\mathcal{K}_T$ denotes the target motif-keyword set extracted from the textual prompt T , whereas $\mathcal{K}_P$ denotes the predicted motif-keyword set derived from the BLIP-generated caption. P_{content} represents the proportion of predicted motif keywords that matched the target motif-keyword set, whereas R_{content} represents the proportion of target motif keywords retained in the candidate synthetic image. The corresponding F1 score was calculated as follows:

$$F_{1,\text{content}} = \frac{2P_{\text{content}}R_{\text{content}}}{P_{\text{content}}+R_{\text{content}}} \quad (4)$$

The local motif-level component primarily evaluated the identity and presence of the target motif. The specified motif quantity was incorporated into the global prompt-image alignment through the natural-language quantity term included in the complete textual prompt, rather than through an independent instance-counting procedure.

The content semantic consistency score was calculated by combining global prompt-image alignment and local motif-level consistency as follows:

$$S_{\text{content}} = \alpha S_{\text{content}}^{\text{global}} + (1 - \alpha)F_{1,\text{content}}, \quad \alpha \in [0,1] \quad (5)$$

α controls the relative contributions of global prompt-image alignment and local motif-level consistency. In this study, α was set to 0.5, assigning equal weights to the two components. $S_{\text{content}}^{\text{global}}$ denotes the global semantic correspondence between the candidate synthetic image and the complete textual prompt, whereas $F_{1,\text{content}}$ denotes the local motif-level consistency derived from keyword matching. A higher S_{content} value indicates stronger overall correspondence with the target content specified in the prompt.

Cultural semantic consistency assessed the extent to which a candidate synthetic image retained the heritage-specific visual meanings represented in its corresponding reference image. These meanings were operationalised as visually identifiable iconographic subjects, symbolic imagery, culturally associated pictorial elements and ornamental motifs. The indicator was quantified by comparing the set of cultural elements identified in the candidate image with the reference-specific cultural-element set derived from a controlled vocabulary.

The controlled vocabulary was constructed from the Cave 38 image corpus, relevant iconographic literature and expert review of the 24 reference images. Following the distinction between iconographic imagery, culturally associated pictorial elements and ornamentation, the initial vocabulary included three categories. Iconographic imagery comprised Buddhist cosmological and symbolic subjects, such as Mount Sumeru, lotus imagery, Bodhi trees and stupas. Culturally associated pictorial elements included paired auspicious birds, guardian lions and stylised sacred mountains. Ornamental motifs included diamond-shaped narrative compartments, decorative lotus patterns, vine scrolls and rosettes. Only elements that were visually identifiable and confirmed as present in at least one reference image were retained in the operational vocabulary. The master vocabulary was defined as

$$\mathcal{V}_{\text{cult}} = \mathcal{V}_{\text{icon}} \cup \mathcal{V}_{\text{pict}} \cup \mathcal{V}_{\text{orn}} \quad (6)$$

The master vocabulary was not treated as a list of elements that every image was required to contain. For each reference image R_i , domain experts defined an image-specific active set containing only the cultural elements visibly present in that reference:

$$C_{R_i} = \{c_k \in \mathcal{V}_{\text{cult}} \mid c_k \text{ is visibly present and expert-confirmed in } R_i\} \quad (7)$$

Each cultural element c_k was expressed as a heritage-qualified visual descriptor T_k , such as “an ancient Kizil mural containing a Buddhist lotus motif” or “an ancient Kizil mural organised as a diamond-shaped Buddhist narrative compartment”. The same pretrained CLIP image and text encoders used for prompt-image alignment were applied in a zero-shot multi-label setting. The normalised element score and the cultural-element set predicted for candidate image I_i were

$$\begin{aligned} s_k(I_i) &= \frac{1 + \cos(f_{\text{img}}(I_i), f_{\text{text}}(T_k))}{2}, \\ C_{I_i} &= \{c_k \in \mathcal{V}_{\text{cult}} \mid s_k(I_i) \geq \tau_{\text{cult}}\} \end{aligned} \quad (8)$$

The threshold τ_{cult} was calibrated once by maximising macro-F1 agreement between the zero-shot CLIP predictions and the expert annotations of the 24 reference images, and was then held

fixed for all candidate synthetic images. Cultural precision and recall were calculated from the overlap between the predicted and reference-specific sets:

$$P_{\text{cult}} = \frac{|C_{I_i} \cap C_{R_i}|}{|C_{I_i}| + \varepsilon}, \quad R_{\text{cult}} = \frac{|C_{I_i} \cap C_{R_i}|}{|C_{R_i}| + \varepsilon} \quad (9)$$

The cultural semantic consistency score was defined as their F1 score:

$$S_{\text{cult}} = F_{1,\text{cult}} = \frac{2P_{\text{cult}}R_{\text{cult}}}{P_{\text{cult}} + R_{\text{cult}} + \varepsilon} \quad (10)$$

Here, $\varepsilon = 10^{-6}$ prevents division by zero. A higher S_{cult} value indicates stronger preservation of the cultural imagery associated with the corresponding reference image. Generic motif identity remained part of content semantic consistency: for example, “two birds” was evaluated as prompted content, whereas “paired auspicious birds within a Buddhist narrative compartment” was treated as a cultural descriptor only when supported by the reference image and expert annotation.

Compositional indicators

The compositional criterion assessed the texture organisation and spatial distribution of visual information in each candidate synthetic image. It comprised three computational indicators: texture characteristics, visual balance and centroid position.

Texture characteristics were evaluated by measuring the correspondence between each candidate synthetic image and its corresponding reference image rather than by assuming that greater contrast, energy or entropy necessarily indicated better mural texture. Both the candidate image I_i and its reference image R_i were converted to greyscale and quantised to 32 grey levels. Symmetric grey-level co-occurrence matrices (GLCMs) were calculated at a one-pixel offset in four directions (0° , 45° , 90° and 135°), and the resulting direction-specific matrices were averaged. Let $G_{a,b}(X)$ denote the averaged co-occurrence count for grey levels a and b in image X . The corresponding normalised probability was

$$P_{a,b}(X) = \frac{G_{a,b}(X)}{\sum_{a,b} G_{a,b}(X)} \quad (11)$$

Contrast quantified the magnitude of local grey-level variation:

$$C(X) = \sum_{a,b} (a - b)^2 P_{a,b}(X) \quad (12)$$

Energy quantified the concentration and regularity of the co-occurrence distribution:

$$E(X) = \sum_{a,b} P_{a,b}(X)^2 \quad (13)$$

Entropy quantified the uncertainty and complexity of the co-occurrence distribution:

$$H_{\text{tex}}(X) = - \sum_{a,b} P_{a,b}(X) \log(P_{a,b}(X) + \varepsilon) \quad (14)$$

For each image X , the texture descriptor vector was $t(X) = [C(X), E(X), H_{\text{tex}}(X)]$. Because the three descriptors have different numerical ranges, their paired differences were standardised using the descriptor-wise standard deviations calculated across the 24 reference images. Let $\sigma_k^{(R)}$ denote the reference-set standard deviation of descriptor $k \in \{C, E, H_{\text{tex}}\}$. The reference-relative texture distance and the corresponding texture-characteristics score were defined as

$$\begin{aligned} d_{\text{texture}}(I_i, R_i) &= \frac{1}{3} \sum_{k \in \{C, E, H_{\text{tex}}\}} \frac{|t_k(I_i) - t_k(R_i)|}{\sigma_k^{(R)} + \varepsilon}, \\ S_{\text{texture}}(I_i, R_i) &= \exp[-d_{\text{texture}}(I_i, R_i)] \end{aligned} \quad (15)$$

Here, $\varepsilon = 10^{-6}$ prevents division by zero. The resulting score lies in the interval $(0,1]$, with a value closer to 1 indicating that the candidate image has GLCM texture statistics closer to those of

its corresponding reference image. Thus, the indicator measures reference-relative texture correspondence and does not interpret an absolute increase in contrast, energy or entropy as intrinsically preferable.

Visual balance was quantified using the normalised Shannon entropy of the spatial distribution of regional image information. Each image was divided into N subregions, and an information measure E_k was calculated for the k th subregion. The proportional contribution of each subregion was then calculated as follows:

$$p_k = \frac{E_k}{\sum_{k=1}^N E_k} \quad (16)$$

E_k denotes the information measure calculated for the k th subregion, N is the total number of subregions, and p_k represents the proportion of the total image information contributed by that subregion. The Shannon entropy of the regional information distribution was calculated as follows:

$$S_{\text{balance}} = -\frac{\sum_{k=1}^N p_k \log p_k}{\log N} \quad (17)$$

A higher S_{balance} value indicated a more even spatial distribution of image information, whereas a lower value indicated that the information was concentrated in fewer subregions. The resulting score was therefore used as a computational proxy for visual balance.

Centroid position was used to quantify the displacement of the image's intensity-weighted visual centroid from the geometric centre of the canvas. Each candidate synthetic image was converted to greyscale. Let $I(x, y)$ denote the pixel intensity at coordinates (x, y) . The coordinates of the intensity-weighted centroid were calculated as follows:

$$x_c = \frac{\sum_{x,y} x I(x, y)}{\sum_{x,y} I(x, y)} \quad (18)$$

$$y_c = \frac{\sum_{x,y} y I(x, y)}{\sum_{x,y} I(x, y)} \quad (19)$$

$\sum_{x,y} x \cdot I(x, y)$ and $\sum_{x,y} y \cdot I(x, y)$ denote the intensity-weighted sums of the horizontal and vertical coordinates, respectively, while $\sum_{x,y} I(x, y)$ denotes the total image intensity. The Euclidean distance between the intensity-weighted visual centroid (x_c, y_c) and the geometric centre (x_0, y_0) was calculated as follows:

$$D_{\text{centroid}} = \sqrt{(x_c - x_0)^2 + (y_c - y_0)^2} \quad (20)$$

(x_0, y_0) denotes the geometric centre of an image with width W and height H , where $x_0 = (W - 1)/2$ and $y_0 = (H - 1)/2$. To make the centroid-position score comparable across images, the distance was normalised by D_{max} , defined as the maximum possible distance from the geometric centre to an image corner:

$$D_{\text{max}} = \sqrt{x_0^2 + y_0^2} \quad (21)$$

$$S_{\text{centroid}} = 1 - \frac{D_{\text{centroid}}}{D_{\text{max}}} \quad (22)$$

The resulting centroid-position score ranged from 0 to 1. A value closer to 1 indicated that the intensity-weighted visual centroid was closer to the geometric centre, whereas a value closer to 0 indicated greater centroid displacement.

Color indicators

Color is one of the most distinctive and expressive visual elements in the Kizil Grotto murals. Its unique system of mineral pigments—such as lapis lazuli blue, atacamite green, and hematite red—not only conveys the symbolic meaning of religious narratives but also reflects traces of painting techniques and cultural exchange. For candidate synthetic Kizil Grotto mural images, if the

generated results fail to accurately reproduce or preserve these characteristic color relationships—even when they perform well in terms of composition and line—the credibility of their overall artistic style may be significantly weakened. Accordingly, the color criterion comprised three computational indicators: color gamut distribution, color contrast and dominant hue consistency.

Color gamut distribution assessed the correspondence between the color distributions of a candidate synthetic image and its corresponding reference image. The candidate image I and its corresponding reference image R were converted to the HSV color space, from which their color histograms $\mathbf{h}_{HSV}(I)$ and $\mathbf{h}_{HSV}(R)$ were extracted.

$$S_{\text{gamut}} = \frac{\mathbf{h}_{HSV}(I) \cdot \mathbf{h}_{HSV}(R)}{\|\mathbf{h}_{HSV}(I)\| \|\mathbf{h}_{HSV}(R)\|} \quad (23)$$

$\mathbf{h}_{HSV}(I)$ and $\mathbf{h}_{HSV}(R)$ denote the color histograms of the candidate synthetic image I and its corresponding reference image R , respectively.

Color contrast was quantified using the mean CIE76 color difference between sampled pixel pairs in each candidate synthetic image. Each image was converted to the CIELAB color space, in which L^* , a^* and b^* represent lightness, the green–red component and the blue–yellow component, respectively. The CIE76 color differences were normalised to the interval $[0,1]$, and the color-contrast score was calculated as follows:

$$S_{\text{contrast}} = \frac{1}{M} \sum_{(i,j) \in \Omega} \Delta E_{ij} \quad (24)$$

ΔE_{ij} denotes the normalised CIE76 color difference between the sampled pixels i and j , Ω denotes the set of sampled pixel pairs, and $M = |\Omega|$ denotes the total number of sampled pairs. A higher S_{contrast} value indicated greater average perceptual color variation within the candidate synthetic image.

Dominant hue consistency assessed the correspondence between the dominant hue of each candidate synthetic image and that of its corresponding reference image. The candidate image and reference image were converted to the HSV color space, and their dominant hue angles were extracted. Because hue is a circular variable, the dominant hue-consistency score was calculated as follows:

$$S_{\text{hue}} = 1 - \frac{\min(|h_I - h_R|, 360 - |h_I - h_R|)}{180} \quad (25)$$

h_I and h_R denote the dominant hue angles of the candidate synthetic image I and its corresponding reference image R , respectively. The score ranges from 0 to 1, with a value closer to 1 indicating greater dominant-hue consistency.

Line indicators

The line criterion assessed two measurable properties of linework in each candidate synthetic image: line smoothness and line-width consistency. Lines in Kizil Grotto murals are characterised by rounded and flowing contours with relatively limited variation in width. Accordingly, line smoothness quantified changes in curvature along extracted line segments, whereas line-width consistency quantified variation in line thickness.

Before calculating the line indicators, line regions were extracted from each candidate synthetic image using a heritage-informed color-aware line-region extraction pipeline developed in this study. The extracted line maps were subsequently skeletonised and divided into individual connected line segments $\mathcal{C}_i$ for curvature and line-width analysis.

Because the contour lines in the experimental images were specified as hematite red, each image was converted from RGB to CIELAB color space, and the color difference between each pixel and the reference hematite-red prototypes was calculated using the CIEDE2000 formulation³⁴. Pixels with a color difference below the reference-calibrated threshold (τ_{red}) were classified as candidate line pixels. Small isolated components were removed and short discontinuities were closed using fixed morphological parameters, producing a binary line-region mask (M).

Line smoothness was quantified using the rate of curvature change along each extracted line segment. For the i th line segment $\mathcal{C}_i$, the segment-level line-smoothness score was calculated as follows:

$$S_{\text{smooth}}(\mathcal{C}_i) = \text{clip}\left(1 - \frac{TV_{\kappa}(\mathcal{C}_i) - a}{b - a}, 0, 1\right) \quad (26)$$

The parameters a and b denote the lower and upper calibration thresholds for the curvature-variation measure, respectively. The linear transformation and clipping operation restrict the resulting score to the interval $[0,1]$, with a value closer to 1 indicating greater line smoothness. $TV_{\kappa}(\mathcal{C}_i)$ denotes the arc-length-normalised total variation of curvature along the i th line segment.

$$TV_{\kappa}(\mathcal{C}_i) = \frac{1}{L_i} \sum_j |\kappa'_j| \Delta s_j \quad (27)$$

$\mathcal{C}_i$ denotes the i th extracted line segment, L_i denotes its total arc length, Δs_j denotes the arc-length interval between adjacent sample points, and κ'_j denotes the rate of curvature change with respect to arc length. A smaller $TV_{\kappa}(\mathcal{C}_i)$ value indicates less curvature fluctuation and therefore greater line smoothness.

The curvature-change rate at sample point j was approximated using the central finite-difference method:

$$\kappa'_j \approx \frac{\kappa_{j+1} - \kappa_{j-1}}{s_{j+1} - s_{j-1}} \quad (28)$$

s_j denotes the arc-length coordinate of sample point j , and κ_j denotes the curvature at that point. The curvature was calculated from the first and second derivatives of the line coordinates as follows:

$$\kappa_j = \frac{|x'_j y''_j - y'_j x''_j|}{\left((x'_j)^2 + (y'_j)^2\right)^{3/2} + \epsilon} \quad (29)$$

x_j and y_j denote the coordinates of sample point j , the prime symbols denote derivatives with respect to arc length, and $\epsilon = 10^{-6}$ is used to prevent division by zero.

Line-width consistency quantified width variation within individual line segments and across the extracted linework of each candidate synthetic image. Specifically, the coefficient of variation, calculated from the mean and standard deviation of local line widths, was used to quantify line-width uniformity:

$$S_{\text{width}}(\mathcal{C}_i) = \text{clip}\left(1 - \frac{CV_i - c}{d - c}, 0, 1\right) \quad (30)$$

The parameters c and d denote the lower and upper calibration thresholds for the line-width coefficient of variation, respectively. The resulting score was restricted to the interval $[0,1]$, with a value closer to 1 indicating greater line-width consistency.

The coefficient of variation for the i th line segment was calculated as follows:

$$CV_i = \frac{\sigma_i}{\mu_i + \epsilon} \quad (31)$$

μ_i and σ_i denote the mean and standard deviation, respectively, of the local line widths measured along the i th line segment. The constant $\epsilon = 10^{-6}$ was used to prevent division by zero.

For each skeleton point p , the local line width was estimated using the Euclidean distance transform:

$$w(p) = 2D(p), \quad p \in S_i \quad (32)$$

$D(p)$ denotes the Euclidean distance from skeleton point p to the nearest boundary of the corresponding binary line region, and S_i denotes the set of skeleton points belonging to the i th line segment.

2.6 AHP weighting and score aggregation

The proposed heritage-aware evaluation framework comprised four criteria—semantics, composition, color and line—and ten corresponding computational indicators. Because these criteria represent different aspects of the suitability of candidate synthetic Kizil Grotto mural images, they were not assumed to contribute equally to the overall evaluation. The Analytic Hierarchy Process (AHP) was therefore used to derive criterion-level weights through structured pairwise comparisons. AHP was applied only at the criterion level, whereas the indicators within each criterion were assigned equal weights. The resulting criterion weights were subsequently combined with the corresponding criterion scores to calculate the overall model scores. The AHP and score-aggregation procedure comprised the following steps:

- (1) The evaluation hierarchy was established with an objective layer, a criterion layer and an indicator layer. The objective layer represented the heritage-aware evaluation of candidate synthetic Kizil Grotto mural images; the criterion layer comprised semantics, composition, color and line; and the indicator layer comprised the ten computational indicators described in Section 2.5.
- (2) A panel of 20 experts, comprising specialists in Kizil mural research, cultural-heritage digitisation and design, was convened to incorporate domain-specific, technical and practice-based perspectives into the criterion-weighting process. The experts independently compared the relative importance of the four criteria for assessing the suitability of candidate synthetic images for inclusion in a Kizil mural training corpus. Each pairwise comparison was conducted using Saaty's 1–9 scale³⁵, as shown in Table 4.

Table 4 | AHP 1–9 scale

Scale	Meaning
1	The two criteria are of equal importance.
3	The former criterion is moderately more important than the latter.
5	The former criterion is strongly more important than the latter.
7	The former criterion is very strongly more important than the latter.
9	The former criterion is extremely more important than the latter.
2, 4, 6, 8	Intermediate values between two adjacent judgements.
Reciprocals	If the pairwise comparison of criterion i with criterion j yields a_{ij} , then the reverse comparison yields $a_{ji} = 1/a_{ij}$.

For n criteria, the pairwise comparisons formed an $n \times n$ reciprocal judgement matrix A , expressed as follows:

$$A = (a_{ij})_{n \times n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \quad (33)$$

where a_{ij} denotes the relative importance of criterion i compared with criterion j , with $a_{ii} = 1$ and $a_{ji} = 1/a_{ij}$.

The individual expert judgements were aggregated using the geometric mean to construct the group judgement matrix. Each expert's judgement was assigned equal importance.

- (3) The principal eigenvector of the group judgement matrix was calculated and normalised to obtain the criterion-weight vector. The resulting weight vector was expressed as

$$W = (w_1, w_2, w_3, w_4)^T \quad (34)$$

where w_i denotes the weight of criterion i , subject to

$$\sum_{i=1}^4 w_i = 1 \quad (35)$$

The four elements of W represented the relative weights assigned to semantics, composition, color and line, respectively.

- (4) The consistency of the group judgement matrix was evaluated using the consistency index (CI) and consistency ratio (CR):

$$CI = \frac{\lambda_{\max} - n}{n - 1} \quad (36)$$

$$CR = \frac{CI}{RI} \quad (37)$$

where $\lambda_{\max}$ is the maximum eigenvalue of matrix A , n is the number of criteria, and RI is the corresponding random consistency index. The group judgement matrix was considered acceptably consistent when $CR < 0.10$.

The experts participated only in determining the relative importance of the four criteria and did not directly assess or score the generated images. All candidate synthetic training images were evaluated using the computational indicators described in Section 2.5.

- (5) Because AHP was applied only at the criterion level, the normalised indicators within each criterion were equally weighted and aggregated using the arithmetic mean. For candidate image I , the four criterion scores were calculated as follows:

$$C_{\text{sem}}(I) = \frac{S_{\text{content}}(I) + S_{\text{cult}}(I)}{2} \quad (38)$$

$$C_{\text{comp}}(I) = \frac{S_{\text{texture}}(I) + S_{\text{balance}}(I) + S_{\text{centroid}}(I)}{3} \quad (39)$$

$$C_{\text{color}}(I) = \frac{S_{\text{gamut}}(I) + S_{\text{contrast}}(I) + S_{\text{hue}}(I)}{3} \quad (40)$$

$$C_{\text{line}}(I) = \frac{S_{\text{smooth}}(I) + S_{\text{width}}(I)}{2} \quad (41)$$

The overall heritage-aware score of candidate image I was then calculated as

$$H(I) = w_1 C_{\text{sem}}(I) + w_2 C_{\text{comp}}(I) + w_3 C_{\text{color}}(I) + w_4 C_{\text{line}}(I) \quad (42)$$

The resulting score ranged from 0 to 1, with a higher value indicating stronger overall correspondence with the semantic, compositional, chromatic and linear requirements of the proposed framework.

For each generative model m , the model-level indicator scores were obtained by averaging the corresponding indicator scores across all valid outputs generated by that model. The model-level semantic, composition, color and line scores were then calculated by equally averaging the indicators assigned to each criterion. The overall model score was calculated as follows:

$$H_m = w_1 C_{\text{sem},m} + w_2 C_{\text{comp},m} + w_3 C_{\text{color},m} + w_4 C_{\text{line},m} \quad (43)$$

The aggregate model score was used for comparative evaluation, whereas the individual indicator scores were retained to identify specific weaknesses that might be obscured by compensatory performance across criteria.

3. Results

3.1 Heritage-aware indicator system

The resulting heritage-aware indicator system followed a three-level hierarchical structure comprising an objective layer, a criterion layer and an indicator layer, as shown in Fig. 3. The

objective layer defined the overall task as the heritage-aware evaluation of synthetic Kizil Grotto mural images intended for use as candidate training assets. The criterion layer consisted of four principal dimensions—semantics, composition, color and line—which represented the major cultural, structural and visual characteristics of Kizil Grotto mural art. The indicator layer operationalised these dimensions into ten measurable indicators.

The semantic dimension comprised content semantic consistency and cultural semantic consistency. Content semantic consistency assessed whether the generated image contained the specified target motif and the required number of motif instances, whereas cultural semantic consistency evaluated its correspondence with the cultural attributes and visual identity of Kizil Grotto murals. The compositional dimension comprised texture characteristics, visual balance and centroid position, reflecting the spatial organisation, distribution of visual information and placement of the principal motif. The color dimension included color gamut distribution, color contrast and dominant hue consistency, which were used to assess the correspondence between generated outputs and the mineral-pigment-based color characteristics of Kizil Grotto murals. The line dimension consisted of line smoothness and line-width consistency, capturing the continuity, regularity and stability of contour-based modelling.

The resulting system translated the heritage-specific characteristics of Kizil Grotto murals into measurable computational variables while maintaining interpretability and comparability across different generative models. It provided the structural basis for subsequent AHP weighting and quantitative evaluation of candidate synthetic training images before their potential inclusion in a training corpus.

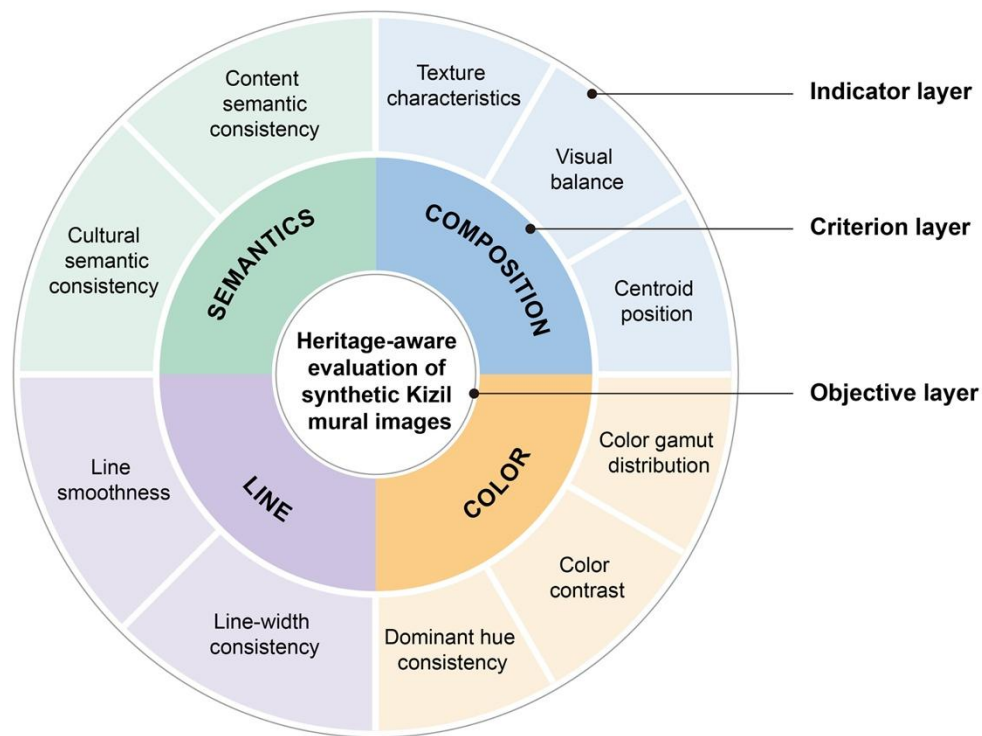


Fig. 3 | Heritage-aware indicator system for synthetic Kizil Grotto mural training images.

3.2 AHP-derived criterion weights

The AHP analysis produced the following criterion-level weights: semantics ($w_1 = 0.3982$), composition ($w_2 = 0.2804$), color ($w_3 = 0.2019$), and line ($w_4 = 0.1195$). The judgement matrix satisfied the AHP consistency requirement ($CR < 0.10$), indicating an acceptable level of consistency in the expert comparisons.

From the overall distribution, the relative weights of the four criteria follow a hierarchical pattern of “Semantics > Composition > Color > Line,” which effectively reflects the artistic characteristics and research objectives of the Kizil Grotto murals. Semantics received the highest weight, accounting for 39.82% of the overall evaluation. This result indicates that motif recognisability and cultural semantic correspondence were prioritised when assessing the suitability

of synthetic Kizil Grotto mural images as candidate training assets. Composition ranked second, with a weight of 28.04%, reflecting the importance assigned to texture organisation, visual balance and the spatial placement of the principal motif. Color accounted for 20.19% of the total weight. It captured the correspondence between generated images and the characteristic chromatic relationships associated with the mineral-pigment palette of Kizil Grotto murals. Line received a weight of 11.95% and represented the continuity, smoothness and width regularity of contour-based modelling. Although color and line received lower weights than semantics and composition, both remained necessary for identifying chromatic and local stylistic deviations in candidate synthetic training images.

3.3 Quantitative comparison of generative models

Table 5 | Quantitative comparison of generative models

Model	Semantics		Composition			Color			Line		Overall Score
	Content	Cultural	Texture	Balance	Centroid	Gamut	Contrast	Dominant	Smoothness	Width	
Stable Diffusion	0.72	0.55	0.68	0.65	0.66	0.70	0.68	0.78	0.62	0.60	0.657
Midjourney	0.80	0.70	0.58	0.72	0.70	0.75	0.69	0.75	0.68	0.66	0.713
DALL·E 3	0.75	0.60	0.70	0.71	0.69	0.73	0.70	0.72	0.66	0.65	0.688
Kandinsky	0.60	0.40	0.62	0.73	0.55	0.60	0.58	0.60	0.65	0.62	0.572

Abbreviations: Content = Content Semantic Consistency; Cultural = Cultural Semantic Consistency; Texture = Texture Characteristics; Balance = Visual Balance; Centroid = Centroid Position; Gamut = Color Gamut Distribution; Contrast = Color Contrast; Dominant = Dominant Hue Consistency; Smoothness = Line Smoothness; Width = Line-Width Consistency.

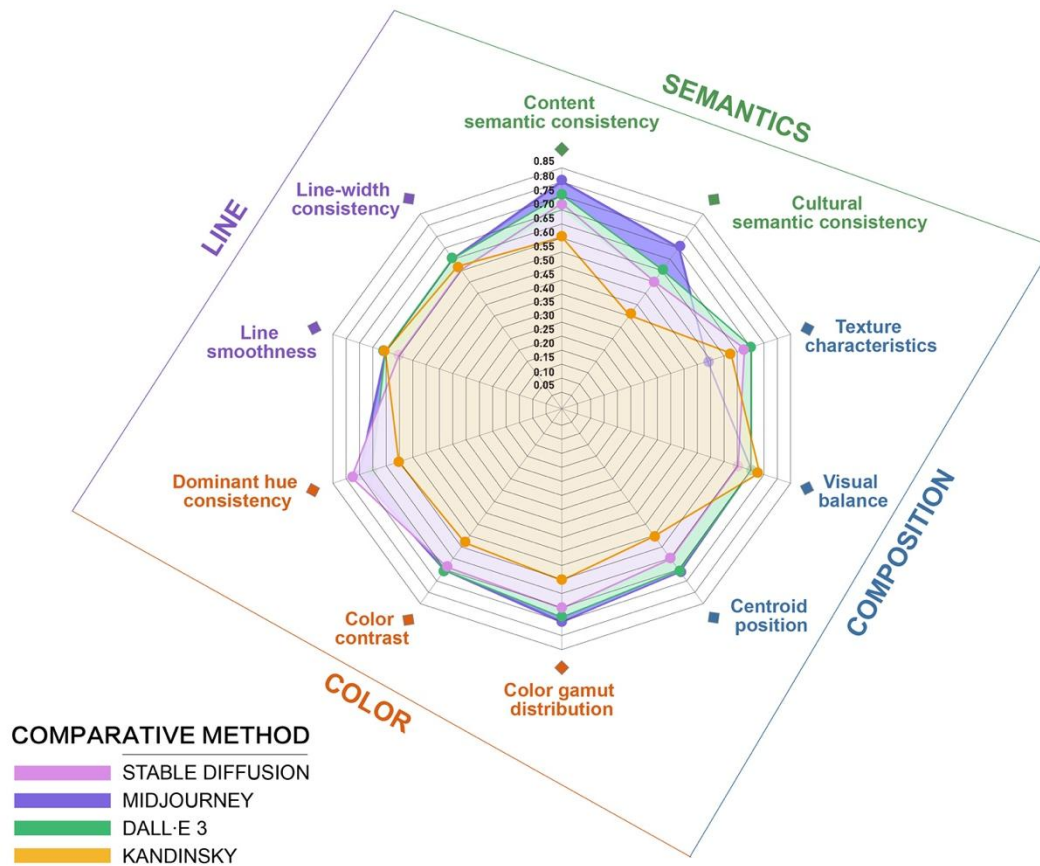


Fig. 4 | Radar chart of model performance across evaluation indicators.

The quantitative comparison was based on 10,119 candidate outputs generated from the same set of 24 reference images depicting non-human motifs. As shown in Table 5 and Fig. 4, the four generative models produced different performance profiles across the ten indicators. Midjourney achieved the highest overall score of 0.713, followed by DALL·E 3 at 0.688, Stable Diffusion at 0.657 and Kandinsky at 0.572.

For the semantic indicators, Midjourney obtained the highest scores for both content semantic consistency and cultural semantic consistency, with values of 0.80 and 0.70, respectively. DALL·E 3 ranked second for both indicators, scoring 0.75 for content semantics and 0.60 for cultural semantics. Stable Diffusion obtained scores of 0.72 and 0.55, whereas Kandinsky recorded the lowest semantic scores, at 0.60 and 0.40.

The compositional results varied across the three indicators. DALL·E 3 achieved the highest score for texture characteristics, at 0.70, and recorded visual balance and centroid position scores of 0.71 and 0.69, respectively. Midjourney obtained the highest centroid position score of 0.70 and a visual balance score of 0.72, but recorded the lowest texture characteristics score of 0.58. Kandinsky obtained the highest visual balance score of 0.73 but the lowest centroid position score of 0.55. Stable Diffusion recorded texture characteristics, visual balance and centroid position scores of 0.68, 0.65 and 0.66, respectively.

No single model achieved the highest score across all three color indicators. Midjourney obtained the highest score for color gamut distribution, at 0.75, DALL·E 3 achieved the highest score for color contrast, at 0.70, and Stable Diffusion recorded the highest dominant hue consistency score of 0.78. Kandinsky obtained the lowest scores for color gamut distribution, color contrast and dominant hue consistency, with values of 0.60, 0.58 and 0.60, respectively.

For the line indicators, Midjourney achieved the highest scores for both line smoothness and line-width consistency, at 0.68 and 0.66, respectively. DALL·E 3 followed closely, with scores of 0.66 and 0.65. Kandinsky obtained scores of 0.65 and 0.62, whereas Stable Diffusion recorded the lowest values, at 0.62 and 0.60.

Overall, Midjourney’s highest aggregate score reflected strong performance in semantic consistency, centroid position, color gamut distribution, line smoothness and line-width consistency, despite its comparatively low texture characteristics score. DALL·E 3 showed a comparatively balanced profile, including the highest texture characteristics and color contrast scores, while Stable Diffusion performed particularly well in dominant hue consistency but obtained lower line scores. Kandinsky achieved the highest visual balance score but received lower semantic, color and overall scores. These results show that the proposed framework can distinguish model-specific performance profiles under the reference-image, prompt and experimental conditions used in this study.

3.4 Qualitative comparison of representative outputs

To complement the quantitative comparison, representative outputs were visually examined in relation to their corresponding reference images and prompt conditions. Fig. 5 presents three motif–color conditions: two birds in atacamite green, a butterfly in lapis lazuli blue, and a lion in gypsum white. Each condition includes outputs generated by Stable Diffusion, Midjourney, DALL·E 3 and Kandinsky. The comparison focused on motif recognisability, internal compositional organisation, correspondence with the specified color category and contour treatment. This analysis was descriptive and did not contribute an additional numerical score. Because the displayed images represent individual outputs, they were used to illustrate image-level variation rather than to validate the aggregate model scores independently.

Across the displayed cases, Midjourney generally preserved the intended motifs and produced centrally organised images. However, the butterfly and lion outputs substantially altered the outer panel shape and local texture organisation of the reference images, consistent with its relatively low texture-characteristics score. DALL·E 3 maintained clear motif recognisability and relatively stable panel structures across the three cases, while presenting cleaner and more regularised visual forms. Stable Diffusion retained the specified dominant color and the overall panel shape in most cases, corresponding to its highest dominant-hue-consistency score. Kandinsky produced visually balanced compositions in the displayed cases but showed greater color variation and more decorative reinterpretation of the reference forms.

The visible differences were interpreted according to the four dimensions of the evaluation framework. Semantic deviations included changes in motif form or weakening of correspondence with the prompted subject. Compositional deviations included alterations in panel shape, motif occupancy, centroid position and local texture organisation. Color deviations were reflected in shifts in hue, contrast or saturation relative to the specified color category. Line deviations included differences in contour smoothness, line-width regularity and the treatment of internal decorative details.

The representative cases broadly reflected the model-level profiles reported in Table 5, although individual outputs did not consistently follow the aggregate trends. Midjourney combined strong semantic, centroid-position and line scores with weaker texture correspondence. DALL·E 3 showed a comparatively balanced performance and the highest texture-characteristics score. Stable Diffusion’s strong dominant-hue consistency was visually apparent in several examples, whereas Kandinsky’s high visual-balance score was accompanied by weaker semantic and color correspondence. These observations indicate that the quantitative results describe general tendencies across the full generated-image dataset, while qualitative inspection reveals image-level deviations relevant to the screening of candidate synthetic training images.

Prompt	Reference image	Stable Diffusion	Midjourney	DALL·E 3	Kandinsky
[two birds], [atacamite green]					

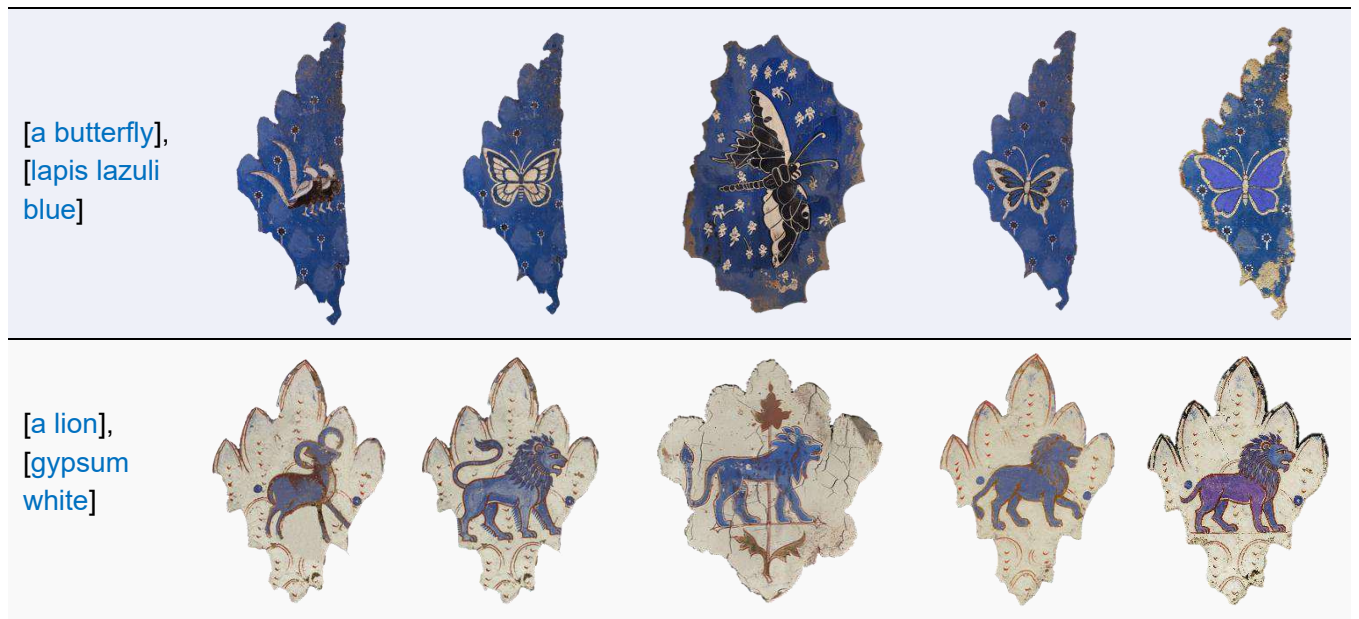


Fig. 5 | Representative outputs across generative models.

4. Discussion

We developed a heritage-aware computational framework for evaluating candidate synthetic Kizil Grotto mural images and supporting their screening before inclusion in training datasets. Under the tested conditions, Midjourney achieved the highest overall score of 0.713, followed by DALL·E 3, Stable Diffusion and Kandinsky. However, the four models exhibited different strengths rather than a uniform performance hierarchy across all indicators. These results should not be interpreted as a universal ranking, because model performance depends on the reference images, prompt conditions, platform settings and model versions. Instead, they demonstrate the need for systematic assessment of synthetic images before their incorporation into a cultural heritage training corpus.

The AHP-derived weights assigned the greatest importance to semantics, followed by composition, color and line. This weighting pattern reflects the importance of motif recognisability, cultural correspondence and pictorial organisation in preserving the visual identity of Kizil Grotto murals. A generated image may retain an appropriate palette or visually regular contours while altering the intended motif or weakening its narrative meaning. Such an image may remain visually attractive but be unsuitable as a training sample. Midjourney's highest aggregate score reflected strong performance in semantic consistency, centroid position, color gamut distribution and the two line indicators, while its score for texture characteristics was comparatively lower. DALL·E 3 showed a comparatively balanced profile and achieved the highest scores for texture characteristics and color contrast. Stable Diffusion obtained the highest dominant-hue-consistency score, whereas Kandinsky achieved the highest visual-balance score but weaker semantic and color correspondence. The representative cases broadly reflected these aggregate profiles, although some individual outputs deviated from the model-level trends. For example, Midjourney preserved recognisable motifs but substantially altered panel shape and local texture organisation in several cases. These variations indicate that aggregate scores represent general model tendencies rather than uniform performance in every output. Qualitative inspection therefore remains necessary for interpreting numerical results and identifying deviations that individual indicators may not fully capture.

Previous mural restoration, inpainting and generation studies have largely focused on model architecture, restoration accuracy, content preservation, artefact reduction, color consistency and visual fidelity¹³⁻¹⁹. General-purpose measures such as SSIM, FID, LPIPS and CLIP-based similarity provide useful assessments of structural, distributional, perceptual and semantic correspondence²³⁻²⁶. Computational aesthetics and image aesthetic assessment have also advanced the quantitative analysis of visual quality, composition and aesthetic attributes³⁶⁻³⁸. However, these approaches do not independently determine whether a synthetic image preserves the combined narrative, compositional, chromatic and linear characteristics of a specific mural tradition. The contribution

of the present framework therefore lies in integrating heritage-specific knowledge into a structured and interpretable quality-control procedure for synthetic training images.

This quality-control function is important because visual plausibility is not equivalent to training-data suitability. Synthetic images may appear coherent while containing motif deformation, compositional alteration, color deviation or irregular contour treatment. If incorporated into training datasets without assessment, such deviations may enter the training corpus and affect downstream model learning. The proposed framework provides quantitative evidence for prioritising candidate images for subsequent expert review and for developing operational retention and exclusion thresholds.

The aggregate score should therefore be interpreted together with the individual indicator scores, particularly those related to semantic consistency, so that strong performance in one indicator does not conceal a critical weakness in another.

Although the heritage-aware evaluation framework proposed in this study provides a structured and interpretable basis for assessing candidate synthetic Kizil Grotto mural images, several limitations remain. First, all source images were derived from Cave 38 and therefore do not represent the full chronological, iconographic and stylistic diversity of the Kizil Grottoes. Second, the generation experiments were restricted to 24 reference images depicting non-human motifs, including animals, plants, landscape motifs and objects. The framework was therefore not evaluated on Buddhist figures, donor portraits or other forms of human iconography represented in Kizil Grotto murals. Third, the manually restored images should not be regarded as definitive reconstructions of the murals' original appearance, and the AHP-derived criterion weights remain dependent on expert judgement. Moreover, the computational indicators were not independently validated against a separate expert-rated image dataset. Fourth, general-purpose vision-language models may not fully recognise heritage-specific motifs and mural attributes, while the continuous updating of closed-source generative systems limits experimental reproducibility. Fifth, the model comparisons were descriptive and were not subjected to inferential significance testing. The operational screening thresholds and downstream training utility of the framework also remain to be validated. Future research should calibrate screening thresholds using independent expert-labelled datasets and compare downstream models trained on real images alone, real images combined with unfiltered synthetic images and real images combined with framework-screened synthetic images. Such experiments would determine whether heritage-aware screening improves restoration, recognition or classification performance. The framework should therefore be understood as a quality-control and decision-support method that complements, rather than replaces, conservation expertise.

Data availability

The datasets analysed during the current study are not publicly available due to institutional and cultural heritage data restrictions, but are available from the corresponding author on reasonable request.

Acknowledgements

This work was supported by the Key Research and Development Program of the Xinjiang Uygur Autonomous Region (No. 2022B03035-1).

Authors' contributions

N.X. conceived the study, developed the evaluation framework and methodology, participated in mural image acquisition and data curation, conducted the experiments and quantitative analyses, prepared the figures, and drafted the manuscript.

X.Z. contributed to the computational methodology, software implementation, data analysis and validation, and reviewed and edited the manuscript.

L.M. provided domain expertise on Kizil Grotto murals and contributed to the heritage-specific interpretation and validation of the results.

H.Z. contributed to the study design and methodology, supervised the research, administered the project, acquired the funding, and reviewed and edited the manuscript.

K.X. administered the project. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

1. Zhao, L., Weng, Z. & Hu, Y. Exploration on digital restoration of Kizil Grotto murals lost overseas (in Chinese). *Journal of Zhejiang University (Science Edition)* **47**, 651–659 (2020).
2. Wu, M., Li, M. & Zhang, Q. Feature separation and fusion to optimise the migration model of mural painting style in tombs. *Appl. Sci.* **14**, 2784 (2024).
3. Wang, J. Study on the diamond-grid landscape images in Kizil Grotto murals (in Chinese). *Decoration* **8**, 136–138 (2022).
4. Zhou, Z. Study on the pigments of Kizil Grotto murals (in Chinese). *Journal of Zhejiang University (Science Edition)* **49**, 726–733 (2022).
5. Yang, Y. Visual art interpretation of the “bent-iron and coiled-silk” line style in Kizil Grotto murals (in Chinese). *Decoration* **12**, 138–139 (2015).
6. Gatys, L. A., Ecker, A. S. & Bethge, M. Image style transfer using convolutional neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2414–2423 (IEEE, 2016).
7. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1125–1134 (IEEE, 2017).
8. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE International Conference on Computer Vision*, 2223–2232 (IEEE, 2017).
9. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695 (IEEE, 2022).
10. Jia, Y. et al. Image inpainting of Thangka murals using edge-assisted feature fusion and self attention based local refine network. *IEEE Access* **11**, 84360–84370 (2023).
11. Sumathi, G. & Uma Devi, M. Inpainting of damaged temple murals using edge- and line-guided diffusion patch GAN. *Front. Artif. Intell.* **7**, 1453847 (2024).
12. Lyu, Q., Zhao, N., Song, J., Yang, Y. & Gong, Y. Mural inpainting via two-stage generative adversarial network. *npj Herit. Sci.* **13**, 188 (2025).
13. Wang, W., Li, Y., Ye, H., Ye, F. & Xu, X. DunhuangGAN: a generative adversarial network for Dunhuang mural art style transfer. In *Proc. 2022 IEEE International Conference on Multimedia and Expo*, 1–6 (IEEE, 2022).
14. Cao, Y., Zhang, Y., Lin, Y. & Wu, K. Dunhuang art style transfer via hierarchical vision transformer and color consistency constraints. *IEEE Trans. Consum. Electron.* **71**, 3240–3251 (2025).
15. Fang, J. et al. Thangka mural style transfer based on progressive style-attentional network and multi-level loss function. *J. Electron. Imaging* **32**, 043007 (2023).

16. Deng, X. & Yu, Y. Ancient mural inpainting via structure information guided two-branch model. *Herit. Sci.* **11**, 131 (2023).
17. Zhao, F., Ren, H., Sun, K. & Zhu, X. GAN-based heterogeneous network for ancient mural restoration. *Herit. Sci.* **12**, 418 (2024).
18. Ren, H., Sun, K., Zhao, F. & Zhu, X. Dunhuang murals image restoration method based on generative adversarial network. *Herit. Sci.* **12**, 39 (2024).
19. Lian, J., Zhang, J., Du, S., Liu, Q. & Liu, J. Adversarial diffusion network for Dunhuang mural inpainting. *IEEE Trans. Circuits Syst. Video Technol.* **35**, 6561–6574 (2025).
20. Sun, D., Mao, X., Shi, J. & Chen, Z. Artistic expression in high-precision digital restoration: a case study of the intelligent restoration of Cave 38 at the Kizil Grottoes in Xinjiang (in Chinese). *Dongfang Shoucang* **11**, 57–59 (2025).
21. Ren, J. Study on the corbelled murals of Cave 38 of the Kizil Grottoes (in Chinese). *Xinjiang Arts* **5**, 11–17 (2022).
22. Miao, L., Tan, L., Xiao, Y. & Pang, J. Investigation report on Caves 38–40 of the Kizil Grottoes in Baicheng County, Xinjiang (in Chinese). *Journal of National Museum of China* **5**, 26–47 (2018).
23. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
24. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proc. Advances in Neural Information Processing Systems* 30, 6626–6637 (Curran Associates, 2017).
25. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 586–595 (IEEE, 2018).
26. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning* **139**, 8748–8763 (PMLR, 2021).
27. Hu, C. et al. Health assessment method of cave murals based on entropy-weighted AHP-cloud model. *npj Herit. Sci.* **13**, 279 (2025).
28. Shi, T., Chen, C., Li, X. & Hao, A. Semantic and style based multiple reference learning for artistic and general image aesthetic assessment. *Neurocomputing* **582**, 127434 (2024).
29. Briellmann, A. A. & Dayan, P. A computational model of aesthetic value. *Psychol. Rev.* **129**, 1319–1337 (2022).
30. Lv, G., Wang, H., Wang, K., Zhao, H. & Zhao, L. Virtual restoration method of Kizil Grotto murals based on multimodal controlled diffusion models. *npj Herit. Sci.* **13**, 554 (2025).
31. Duan, T., Wang, T., Tao, L. & Liu, B. TKSore for generated Thangka image quality assessment with perceptual alignment and structural constraints. *npj Herit. Sci.* **14**, 395 (2026).
32. Elsharif, W., Agus, M., Alzubaidi, M. & She, J. Cultural Relevance Index: measuring cultural relevance in AI-generated images. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 410–416 (IEEE, 2024).

33. Li, J., Li, D., Xiong, C. & Hoi, S. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proc. 39th International Conference on Machine Learning* **162**, 12888–12900 (PMLR, 2022).
34. Sharma, G., Wu, W. & Dalal, E. N. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Res. Appl.* **30**, 21–30 (2005).
35. Saaty, T. L. The analytic hierarchy process—what it is and how it is used. *Math. Model.* **9**, 161–176 (1987).
36. Yang, S. et al. A self-supervised image aesthetic assessment combining masked image modeling and contrastive learning. *J. Vis. Commun. Image Represent.* **101**, 104184 (2024).
37. Bo, Y., Yu, J. & Zhang, K. Computational aesthetics and applications. *Vis. Comput. Ind. Biomed. Art* **1**, 6 (2018).
38. Brachmann, A. & Redies, C. Computational and experimental approaches to visual aesthetics. *Front. Comput. Neurosci.* **11**, 102 (2017).