npj | heritage science

Article

https://doi.org/10.1038/s40494-025-02167-y

Imagest: a style conditioned art painting synthesis and generation networks

Jinghao Hu[1], Pengbo Zhou[2], Jian Gao[1], Yuhe Zhang[1] ✉, Guohua Geng[1] ✉ & Mingquan Zhou[1]

To address the restoration of artworks and enhance the conservation of uniquely styled paintings using computational techniques such as large-parameter models, we introduce a diffusion-based network, namely Imagest, a novel multi-conditional artwork synthesis method. Specifically designed to overcome the limitations of existing restoration techniques for heavily damaged artworks, Imagest addresses the challenges of insufficient conditioning and inconsistent synthesis results. By jointly leveraging stylistic image and text prompts, our method facilitates more accurate and stylistically coherent reconstructions. Experiments conducted on the WikiArt and Chinese traditional painting datasets demonstrate Imagest's efficiency in generating artworks that align with given style and content prompts. Compared to state-of-the-art baselines such as DALL.E 2 and Stable Diffusion, Imagest achieves competitive performance in both image inpainting and text-guided synthesis, as evidenced by favorable FID and CLIP-score metrics.

Art paintings constitute invaluable treasures of cultural heritage, showcasing historical esthetics and techniques, and possessing irreplaceable cultural significance and standing. However, existing paintings have suffered varying degrees of damage, making the restoration and rehabilitation of art paintings a pivotal branch of heritage science. The task of repairing damaged areas within artworks requires meticulous consideration of both the artistic style and the content of the affected area[1–4]. The stylistic cues for the region requiring restoration can be inferred from the adjacent, unharmed portions of the artwork, whereas the content can be suggested by pertinent textual references.

In recent years, the advent of deep generative models has opened new avenues for digital restoration. Among them, methods based on Generative Adversarial Networks (GANs) have shown considerable potential in artwork synthesis[5–9]. Despite their success in image generation tasks, GANs are inherently constrained by their reliance on training datasets, typically producing outputs in a limited stylistic range. Moreover, GANs often suffer from mode collapse[7], leading to reduced diversity in the results. Introducing additional restrictions, such as text and style guidance, can further compromise the quality of restoration output[7–9].

Recently, the diffusion model has shown amazing capabilities as a generative model, which is a Markov chain-based latent variable model and can avoid mode collapse of GANs[10]. Diffusion model-based methodologies have attained state-of-the-art performance in practical applications, exemplified by systems such as DALL.E 2[11] and Imagen[12]. However, these approaches continue to require substantial computational resources and exhibit considerable temporal demands. While the Stable Diffusion model addresses the computational intricacies associated with the Latent Diffusion Model (LDM), a notable deficiency persists in the realm of restoration tasks pertaining to historical artworks, particularly when conditioned on artistic style and textual prompts.

On the other hand, efficiency remains a factor that must be considered. In conservation practice, a large proportion of paintings present minor to moderate losses spread across the surface. Although digital inpainting for a single image can complete within seconds, the end-to-end workflow-damage mapping and mask delineation, high-fidelity digitization, curator-conservator verification, and re-rendering-remains the dominant cost. Recent reports on physically implementing digitally computed infill indicate that, even with AI assistance, a single full restoration step may still require about 3.5 h, albeit markedly shorter than traditional interventions that can extend to days or weeks[13]. These observations motivate our focus on computational efficiency: lowering latency and memory footprint yields tangible benefits when scaled to museum backlogs and enables more frequent expert iterations.

To accomplish image generation tasks conditioned on style and text using a reasonable amount of training data and time, we introduce **Imagest**, a **s**tyle- and **t**ext- conditioned **image** generation network that integrates the Swin Transformer[14] and the LDM[15]. Within Imagest, we devise a specialized Swin Transformer variant, termed the Style Swin Encoder, dedicated to extracting intricate style features. Additionally, we leverage the CLIP model to produce features from text prompts[16]. Subsequently, we align both the style and text features within the latent space of the LDM. In particular, within the Style Swin Encoder, we acknowledge that style features, which

[1]School of Information Science and Technology, Northwest University, Shaanxi, China. [2]School of Arts and Communication, Beijing Normal University, Beijing, China. ✉e-mail: zhangyuhe0601@nwu.edu.cn; ghgeng@nwu.edu.cn

encompass lines, curves, colors, and other attributes, are largely decoupled from content semantics. Hence, we dispense with the intricate window-shifting operations in the standard Swin Transformer. Instead, for the attention mechanism, we adopt a simplified window multi-head attention mechanism, grounded on relative position encoding based on deviation values, thereby mitigating computational overhead[14]. Our proposed network boasts the following advantages:

(i) Reduced Computational Demand: The LDM circumvents intricate computations in the high-dimensional pixel space, while the Style Swin Encoder utilizes a less computationally intensive window attention mechanism for feature extraction.

(ii) Distinctiveness of the enhanced image feature: The hierarchical structure of the Swin Transformer, complemented by the window attention mechanism, facilitates superior extraction of stylized feature details, resulting in images with more prominent and distinctive characteristics.

In summary, the key contributions of this work are as follows.

- Imagest Network: We propose Imagest, a synthesis network utilizing LDM and Swin transformers to generate or restore artwork from CLIP embeddings and style images with minimal training and dual feature integration.
- Styled Swin Encoder: We design a novel Swin encoder with windowed multi-head attention to efficiently extract stylized visual features while reducing computational cost.

Methods

Related work

Artwork restoration encompasses two primary approaches: manual restoration and digital restoration using computer technology for image synthesis. Both methods play crucial roles in preserving the integrity and beauty of art pieces, each with its unique set of tools, techniques, and applications.

Manual restoration, often referred to as traditional restoration, involves hands-on intervention in artworks by highly skilled conservators[17,18] This approach relies heavily on the expert's knowledge of art history, materials science and artistic techniques. Manual restorers use a variety of tools and materials to clean, consolidate and repair damaged artworks, addressing issues such as cracks, tears, fading, and structural weaknesses. Their work requires meticulous attention to detail, as even the smallest mistake can irreversibly alter the artwork's appearance and value. Concurrently, hand restoration exhibits certain limitations, being a methodology that is not only time-consuming and labor-intensive but also has the potential to inflict secondary damage upon the artwork in question.

With the advancement of computer vision, digital restoration methods have gained prominence[19,20]. Early approaches focused on texture synthesis and regression-based techniques to predict missing image regions[21]. These methods were limited by low creativity and insufficient data representation[7].

The introduction of deep learning led to the adoption of GANs[5,6], which improved the restoration quality and adaptability of the data. However, GANs face challenges such as training instability and mode collapse[22]. More recently, large-parameter models and multimodal architectures have opened up new possibilities in restoration tasks[12,15], with vision-language models like CLIP[16] contributing additional flexibility and creativity.

These developments highlight the transition from manual to algorithmic methods in artwork restoration and suggest that multimodal models can play a central role in bridging semantic and stylistic cues.

In parallel, image synthesis has emerged as a powerful tool for generating artistic content, with techniques evolving from GANs to more recent transformer- and diffusion-based models.

Until the advent of diffusion models[11,12,15], GANs held sway over the realm of image synthesis tasks, with text-to-image generation being a pivotal subdomain. Since their inception[7], GANs have produced images that closely resemble real-world imagery due to their reliance on Nash equilibrium and adversarial training, which presents a viable solution to the text-to-image challenge[9,22,23]. Scott Reed et al. were pioneers in leveraging GANs for text-to-image synthesis[8], initiating the process by parsing the input description through natural language processing to instruct the generator in outputting an accurate and natural depiction of the text. The discriminator then describes and discriminates the generated image, participating in an iterative dance with the generator[8]. The text-conditional convolutional GAN marked the debut of image synthesis driven by text, albeit with results that were short of expectations, requiring extensive training to yield a singular outcome[9]. The limitation of producing a single description per result has persisted as a recurring issue in GANs.

Subsequently, Zhang H et al. introduced StackGAN[23] and StackGAN++[9], dividing the synthesis task into a multistage process where the first stage generates coarse results and the second stage refines these details to produce high-quality images. However, this approach introduced computational complexity, rendering training arduous and susceptible to mode collapse[9,22]. In contrast to GANs, Variational Autoencoders (VAEs)[24], a likelihood-based generative model, possess a more refined structure and expedite image synthesis, albeit at the expense of image quality compared to GANs. Furthermore, Auto-regressive (AR) transformers[24,25], which integrate CLIP[16] and GPT[26–28], demonstrate the capability to generate more intricate and superior quality images. However, this comes at the cost of increased computational complexity[25].

Among these advances, diffusion models have gained significant attention due to their robustness and ability to generate high-fidelity images without suffering from issues such as mode collapse Fig. 1.

The Denoising Diffusion Probabilistic Model (DDPM) was introduced and adapted for text-to-image tasks by Ho et al. [10]. Diffusion models, which are generative models grounded in Markov chain principles[29], encompass two primary stages: a diffusion process that transitions from the original

Fig. 1 | An artwork is comprised of both content and artistic style. The combination of content with various styles can provide a continuous stream of imagination for the artwork, and also offers more possibilities for the restoration of artistic pieces.

image to normally distributed noise and a reverse process that reconstructs the original image from the noise[11,12,15,30,31]. In particular, diffusion models are resilient to mode collapse due to their capacity to preserve the semantic structure of the data[10,29].

Ramesh A et al. presented DALL.E 2, a seminal text-to-image model, which leverages the powerful language-image model CLIP[16] and a guidance diffusion model known as GLIDE[32]. DALL.E 2 comprises two components: a prior and a decoder[11]. The prior transforms text embeddings into image embeddings, which are then utilized to modulate the diffusion decoder to produce the final image. Compared to DALL.E 2[11], Imagen[12] enhances image realism by increasing the size of the language model and employs a continuous diffusion model to refine the generated image. By freezing the weights during text embedding, Imagen simplifies complex structures, reduces computational overhead, and achieves state-of-the-art results in image synthesis. However, both DALL.E 2 and Imagen require extensive training times, often measured in hundreds of GPU days. The LDM addresses this challenge by incorporating latent space processing[15].

The core of diffusion-based image generation lies in feature reconstruction from noise through iterative denoising steps. However, the success of this process is contingent upon the model's ability to extract representative features. In this context, transformer architectures-originally developed for NLP-have demonstrated superior capability in visual feature extraction, leading to their widespread adaptation in vision models such as ViT and Swin[33]. Motivated by this success, the field of computer vision has gradually embraced the transformer as a potential replacement for convolutional neural networks. However, initial attempts failed to yield promising results as a result of the transformer's high computational demands. To address this, Google Labs introduced ViT (Visual Transformer)[34], which replaced the traditional pixel-based visual processing unit with image patches. This innovation overcame the significant challenge of computational complexity and paved a new path for image processing tasks, enabling the exploration of novel methodologies[14,35–37].

Despite its potential, ViT necessitates a substantial amount of data and complex computational resources. To mitigate these limitations, Liu et al. proposed the Swin Transformer[14], a novel visual transformer based on a hierarchical feature map. The Swin Transformer employs a shifted window multi-head attention mechanism (SW-MSA) to significantly reduce computational complexity. Furthermore, its hierarchical structure allows for the extraction of more detailed image features, leading to superior performance in several fundamental image processing tasks[14,36,37].

Our work

To achieve the task of generation of style and text-conditioned artwork, we introduce a novel network architecture termed Imagest. This network integrates input text features and style features through a sophisticated cross-attention mechanism and autoregressive processes. The Imagest architecture primarily encompasses two key modules: the LDM module and the feature prompts module. Within the feature prompts module, the text features are derived from a pre-trained CLIP model[16], while the style features are obtained from a style-specific Swin encoder. A detailed illustration of our network architecture is provided in Fig. 2.

Building on the latent space representations established in prior work, we now introduce the core generative backbone of our framework, which governs the transformation from textual and stylistic conditions to synthesized images.

Our LDM module builds upon prior research endeavors, notably[15] and ref. 35, and comprises two fundamental components: a pre-trained autoencoder and a diffusion model. The pretrained autoencoder plays a pivotal role in facilitating perceptual image compression, thereby mitigating the computational complexities associated with high-dimensional spaces. Conversely, the diffusion model undertakes the task of image generation.

The input image is initially passed through the pre-trained autoencoder[35], which then feeds its output into the diffusion model[15]. This sequence of operations maps the image from a high-dimensional pixel space to a more compact, low-dimensional latent space. The autoencoder, denoted as ε, is trained using a blend of perceptual loss and adversarial loss to enhance its performance. Following denoising by the diffusion model, a decoder, designated as D, is employed to reconstruct the latent space features back into a coherent image, effectively reversing the compression process executed by the encoder ε.

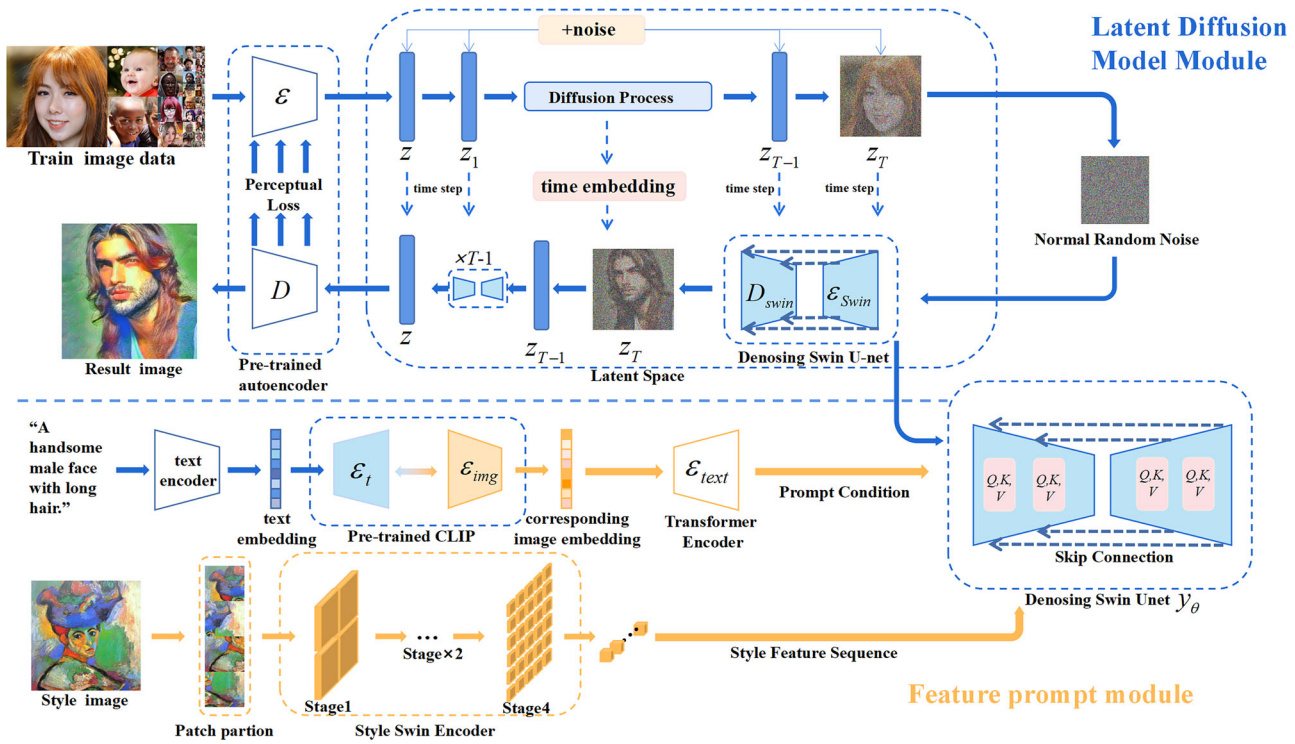

Fig. 2 | The network architecture of Imagest comprises two integral modules. The first module, denoted in blue, is the LDM module. The second module, highlighted in orange, encompasses the style Swin encoder and the CLIP module.

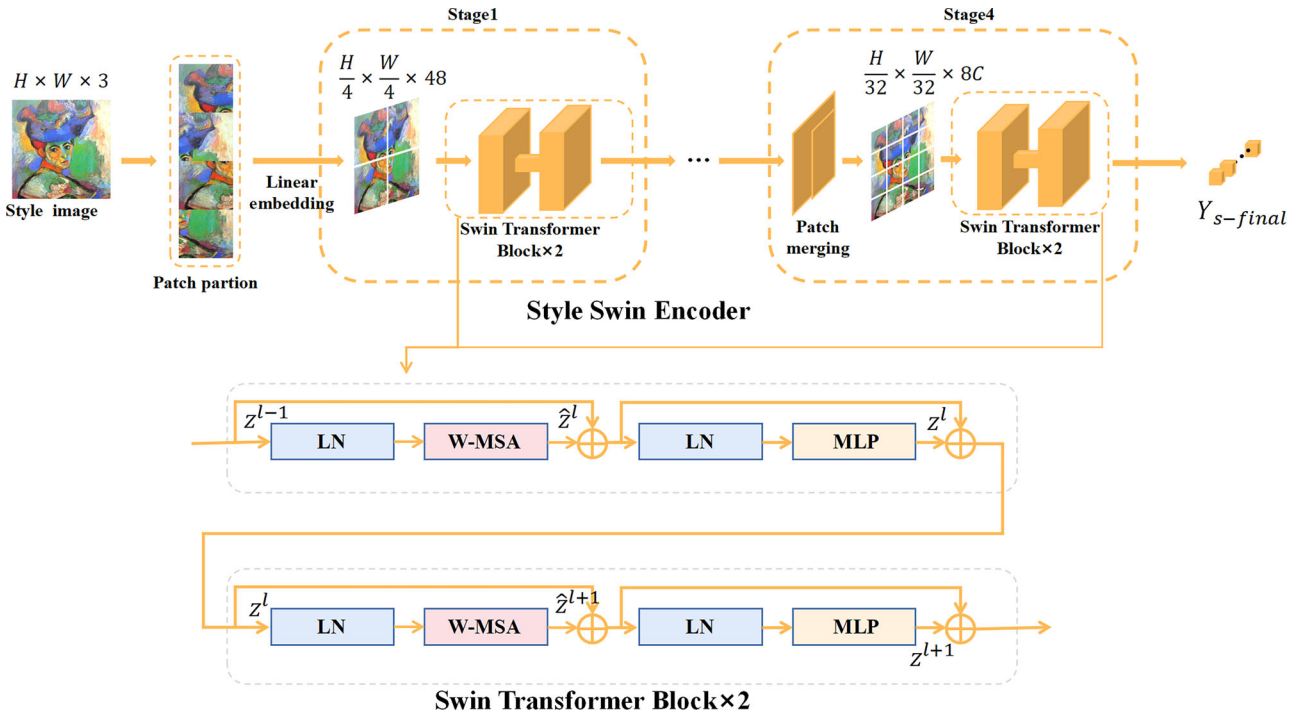

Fig. 3 | The structure of style Swin encoder. The Style Swin Encoder processes patch-partitioned style images through four stages of non-shifted Swin Transformer Blocks, removing the original shifted-window self-attention to reduce computation and avoid unwanted global semantic mixing in style representation.

In the latent space, the diffusion model is segmented into two distinct phases: the diffusion process and the inverse diffusion process, as detailed in prior works[10,29]. The diffusion process involves transforming the training sample features into independent samples that conform to a normal distribution through the iterative addition of noise, facilitated by components such as TimeEmbedding and Residual Block[10,29,38,39]. Subsequently, our proposed inverse diffusion process leverages a denoising U-net block architecture to progressively refine each Gaussian-distributed sample feature, adhering to the predefined noise addition strategy for effective denoising.

In more detail, given an RGB image Image $\in \mathbb{R}^{H\times W\times 3}$, this image first goes to the encoder ε, which will map it to the lower dimensional latent space in $z = \varepsilon(\text{Image})$. Subsequently, z will enter the diffusion process,after T time steps of noise addition, z turns into z^T at step T. z^T are satisfied:

$$z^T = \sqrt{1-\beta_T}z^{T-1} + \sqrt{\beta_T}y_{T-1} \quad (1)$$

with $\beta_T \in [0, 1]$ being a hyperparameter of the diffusion schedule and $y_{T-1} \sim N(0, 1)$ being a standard Gaussian noise. Due to the known noise addition strategy, for each step T, z^T can be calculated iteratively, which will be used to optimize the diffusion model.

Next, the standard Gaussian sample z^T is refined and denoised, and in each denoising step, we use a denoising U-net $y_\theta(T, z^T)$[37], The optimization of each iteration should be:

$$L = \mathbb{E}_{\varepsilon(\text{image}),y_{T-1}\sim N(0,1),T}\left[\left\|y - y_\theta(z^T, T)\right\|_2^2\right] \quad (2)$$

LDM possesses the capability to incorporate conditional constraints, such as image-based and text-based constraints, by leveraging the cross-attention mechanism embedded within the denoising U-Net architecture. Our research focuses on refining and composing images through the fine-tuning of stable diffusion 1.4. Specifically, during this process, we freeze the parameters associated with the noise addition component. Meanwhile, text and style cues are introduced as conditional inputs via the cross-attention mechanism. Additionally, we incorporate fixed-style features by training a style extractor in conjunction with the denoising U-Net. Ultimately, the newly generated image is reconstructed using a pre-trained decoder from latent space, yielding the desired final result.

To effectively incorporate both visual style and textual semantics, we developed a dual-branch conditioning module. The following section outlines the strategies employed to extract and align these conditioning signals.

Given a style image Image $\in \mathbb{R}^{H\times W\times 3}$ and a content text that delineates the specific content requirements, our objective is to utilize the style image to render an image that corresponds to the described text, thereby generating a distinctive artwork. To accomplish this task of style and text-conditioned image generation within a feasible amount of training data and computational time, we initially leverage the pretrained CLIP[16] to generate text embeddings $e_{crt} \in \mathbb{R}^{T_x\times d_t}$, where T_x represents the tokens derived from the content requirements text, and d_t denotes the dimension of the token embeddings. Subsequently, we extract style sequences through our specially designed Style Swin Encoder. The architecture of our Style Swin Encoder is illustrated in Fig. 3.

For capturing style-specific representations, we adopt a lightweight transformer-based encoder. This design balances expressiveness and efficiency, making it well-suited to our application domain.

In most diffusion-based conditional generation frameworks, Vision Transformers (ViT) are commonly used as image encoders due to their global receptive field and strong feature extraction capabilities[16,40]. However, for high-resolution image inputs, ViT incurs substantial computational cost due to its quadratic complexity with respect to spatial dimensions. To address this, we adopt the Swin Transformer[14] as our style encoder.

Swin Transformer introduces a hierarchical structure with shifted window-based attention, which significantly reduces computational complexity while maintaining a strong local feature learning capability. According to the analysis in ref. 14, the theoretical floating point operations (FLOPs) of the Swin Transformer are substantially lower than those of ViT under high-resolution inputs. For example, in our task, when processing a 256 × 256 image:

- ViT-Base (patch size 16 × 16): ~17.6 GFLOPs,
- Swin-Base: ~4.5 GFLOPs.

This represents a reduction of nearly 75% in the computation, making the Swin Transformer more suitable for our application scenario, where efficient but expressive style encoding is essential. Furthermore, we simplify

the original Swin architecture by removing window-shift operations and reducing attention depth, resulting in additional computational savings without significantly affecting representation quality.

To further reduce the computational overhead and tailor the encoder for style representation rather than dense semantic modeling, we simplify the original Swin architecture in two key ways. First, we remove the shifted window mechanism, which is mainly designed to enhance long-range dependency modeling but introduces additional complexity and memory operations. Second, instead of using relative position bias with cyclic shift, we adopt a simpler absolute deviation-based positional encoding scheme. This not only improves training stability but also leads to faster inference with negligible impact on style representation performance.

Our Style Swin Encoder comprises four Swin stages, with each stage containing two Swin Transformer blocks and one Patch Merging block. Since the extracted features are low-dimensional style features, the concern of information exchange between different windows is mitigated, allowing for the adoption of a simpler window attention mechanism. Specifically, an RGB pixel image $\text{Image} \in \mathbb{R}^{H \times W \times 3}$ is fed into the Patch Partition layer, where it is flattened in the channel direction to obtain image patches. The size of each image patch is 4×4, and the feature dimension of the image patch is 48 (resulting from $4 \times 4 \times 3$). Consequently, the dimension of the image is transformed from $H \times W \times 3$ to $\frac{H}{4} \times \frac{W}{4} \times 48$ after passing through the Patch Partition layer. Following this, a linear embedding layer[14] is applied to this initial feature map, projecting it to an arbitrary dimension (denoted as C). As a result, the dimension of the feature map of the image becomes $\frac{H}{4} \times \frac{W}{4} \times C$, and a linear sequence Z_S of feature map styles is obtained.

Subsequently, the sequence Z_S is fed into a sequence of two contiguous Swin-Transformer blocks. A Swin Transformer block primarily comprises a window-based multi-head self-attention mechanism (W-MSA) followed by a multi-layer perceptron (MLP). Notably, a linear normalization layer precedes both the W-MSA and the MLP components. As a result, the combination of the initial linear embedding layer and the subsequent contiguous double Swin Transformer block constitutes the first encoding stage of the Swin Transformer. During this encoding process, the sequences are transformed into query matrix Q, key matrix K, and value matrix V for the computation of the attention mechanism, as follows:

$$Q = Z_S W_Q, \quad K = Z_S W_K, \quad V = Z_S W_V \tag{3}$$

Here, W_Q, W_K, and W_V belong to the space $R^{C \times d_{\text{head}}}$, where $d_{\text{head}} = \frac{C}{N}$, and N denotes the number of heads in the multi-head attention mechanism. The attention mechanism, which incorporates relative position encoding via a bias term $B \in \mathbb{R}^{M^2 \times M^2}$, is calculated as:

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{SoftMax}\left(\frac{\text{QK}^{\text{T}}}{\sqrt{\text{d}}} + \text{B}\right)\text{V} \tag{4}$$

where d is the dimension of Q/K, without considering the issue of information exchange between different windows. Next, the input style sequence is entered into a new Swin Transformer block and computed:

$$\begin{aligned} \hat{z}^l &= WMSA\left(LN\left(z^{l-1}\right)\right) + z^{l-1}, \\ z^l &= MLP\left(LN\left(\hat{z}^l\right)\right) + \hat{z}^l \end{aligned} \tag{5}$$

Consequently, we obtained the final style sequence $Y_{s-final}$ passing through the entire style Swin encoder.

After all adding noise processes, we use cross attention to introduce the style feature to latent space in denoising Unet decoder, precisely, this method generates key matrix(K) and value matrix(V) using style sequences, while Q remains unchanged:

$$Q = Y_C W_Q, K = Y_S W_K, V = Y_S W_V \tag{6}$$

To extract semantic cues from natural language, we leverage a pre-trained vision-language model. This allows textual instructions to guide image synthesis in a highly controllable and interpretable manner.

CLIP stands as a formidable vision-language model, having been established as a central conduit for text-based image generation tasks[16]. While the T5 model, employed within the Imagen framework, enhances the comprehension of textual content, it encounters challenges in reconstructing image features that align seamlessly with the given description. Furthermore, T5 struggles to grasp the semantics that underpin image layouts[40]. Conversely, pre-trained CLIP boasts an extensive data repository. Consequently, we opt for pre-trained CLIP[16] as our prompt condition encoder. In detail, we procure text embeddings through a pre-trained CLIP model, which are subsequently fed into a transformer to obtain a latent encoding. This latent encoding is then mapped to the LDM utilizing the cross-attention mechanism within a designated window.

Results

The proposed Imagest network is capable of successfully synthesizing and repairing images that are both high-quality and highly stylized. Through the implementation of our novel feature cueing module, which efficiently introduces styles corresponding to the source style image, we conduct rigorous experimental validation of our methodology. Specifically, we evaluate the effectiveness of Imagest in two core tasks: the generation of Western- and Chinese-style artworks from textual prompts and the restoration of damaged paintings using style guidance. Both tasks are assessed through qualitative comparisons and quantitative metrics, demonstrating the versatility and performance of our approach.

Implementation details

To evaluate the robustness of Imagest across diverse artistic styles, we employed the Wiki Art dataset[41], encompassing 15 distinct genres of Western art, such as impressionistic and abstract oil paintings, alongside the ChipPhi and TCLPD datasets[42,43], which collectively feature 4168 high-quality Chinese oil paintings, predominantly traditional landscape and ink painting styles.

Our experimental setup was conducted within an Anaconda virtual Python environment on Ubuntu 22.04, utilizing Python version 3.8 and PyTorch version 1.10.1. All experiments were executed on an RTX 4090D GPU equipped with 24GB of video memory. The fine-tuning training comprised 13,600 steps. Furthermore, the autoencoder utilized by our LDM had dimensions of $4 \times 64 \times 64$, and the downsampling factor f was set to 4, aligning with previous studies that have shown that this configuration is optimal for the LDM performance[15].

To ensure fairness and reproducibility in all comparative evaluations, we clarify the following experimental assumptions. First, the results for DALL.E 2 and Imagen were either cited from published literature or obtained through their official APIs and public platforms, as neither model has released a full training code or weights for reproduction. Second, all fine-tuning baselines-DreamBooth, Textual Inversion, and LoRA-were trained using the same datasets (WikiArt, ChipPhi, and TCLPD), with all images uniformly resized to 256×256 pixels during preprocessing. Third, all experiments, including training and evaluation, were performed on a single RTX 4090D GPU (24GB VRAM). The hardware and configuration constraints related to Imagest have been explicitly documented in the "Data availability" section.

Style-guided painting synthesis

For each style, extensive generation experiments totaling hundreds of iterations were conducted, each iteration utilizing distinct textual descriptions. The results are presented in Fig. 4 and Fig. 5. It is evident from the results that each generated image has unique artistic attributes and aligns closely with the input textual prompt. Notably, Imagest also demonstrates the capability to create imaginative works that transcend reality, as exemplified by certain images, such as "A Terra Cotta Warrior playing the piano", "The rocket launches into the sky in Chinese painting". Figure 4 shows the

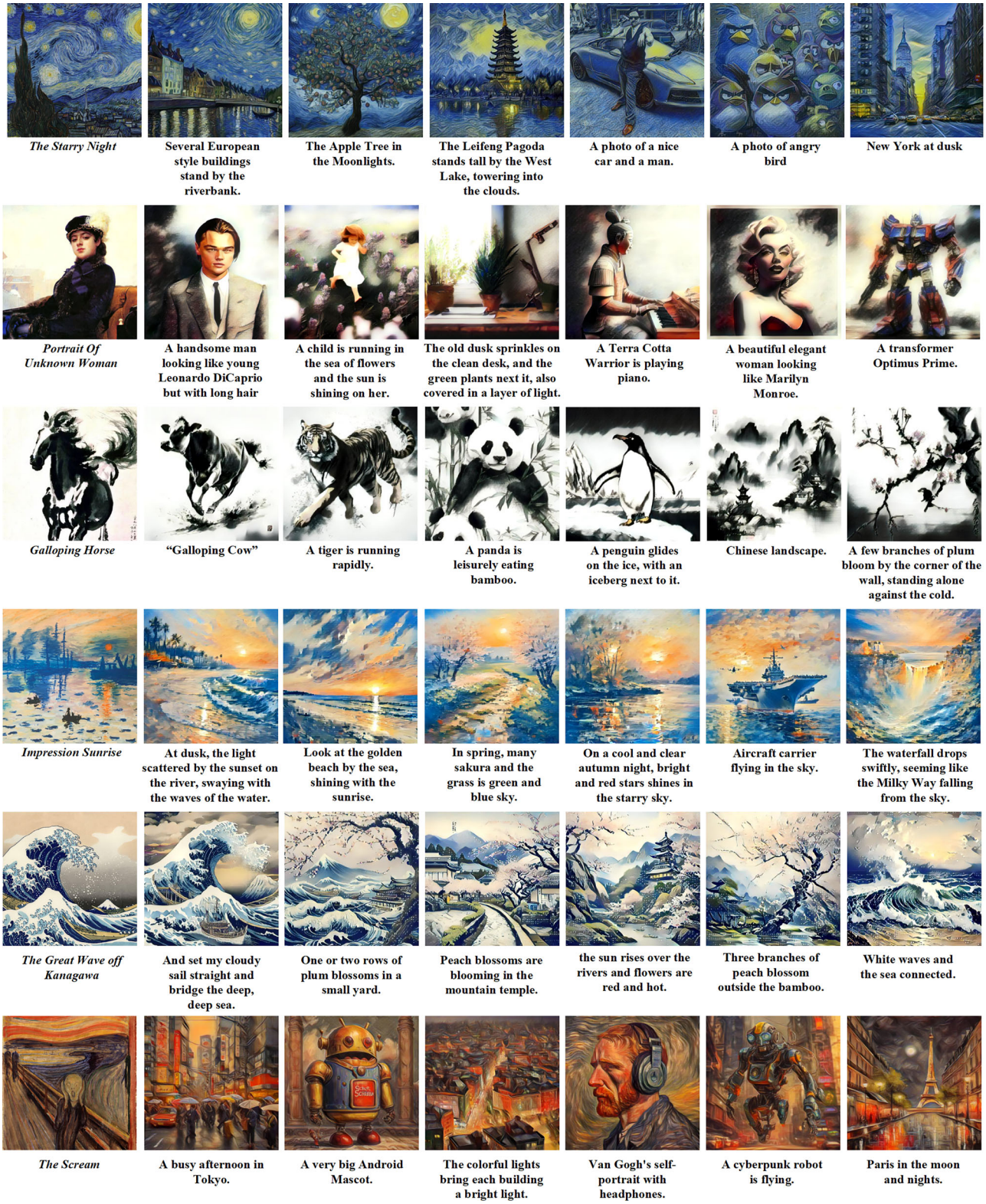

Fig. 4 | The results of Imagest on text and oil style conditioned art painting generation. Each pair shows a style reference (left) and the stylized result (right), covering oil paintings, ink wash, traditional Chinese painting, sketches, and modern styles, demonstrating excellent transfer of texture, color, and strokes while preserving content structure.

resultant artwork synthesized in the manner of an oil painting, whereas Fig. 5 exhibits the synthesized artwork that embodies the style of Chinese painting.

Imagest can perform the task of image generation in almost all styles and generate distinctive artistic images, which is not available in the previous image generation models[11,12,15,24,25], to the best of our knowledge. To verify this observation, we compared Imagest with some current baseline methods: Stable Diffusion[15], DALL.E 2[11], and Imagen[12].

Considering the fact that all the aforementioned models are designed to generate images based on textual input, we initially conducted an experiment to assess the sensitivity of the baseline model to stylistic descriptions within text. The experimental results, presented in Fig. 6, indicate that

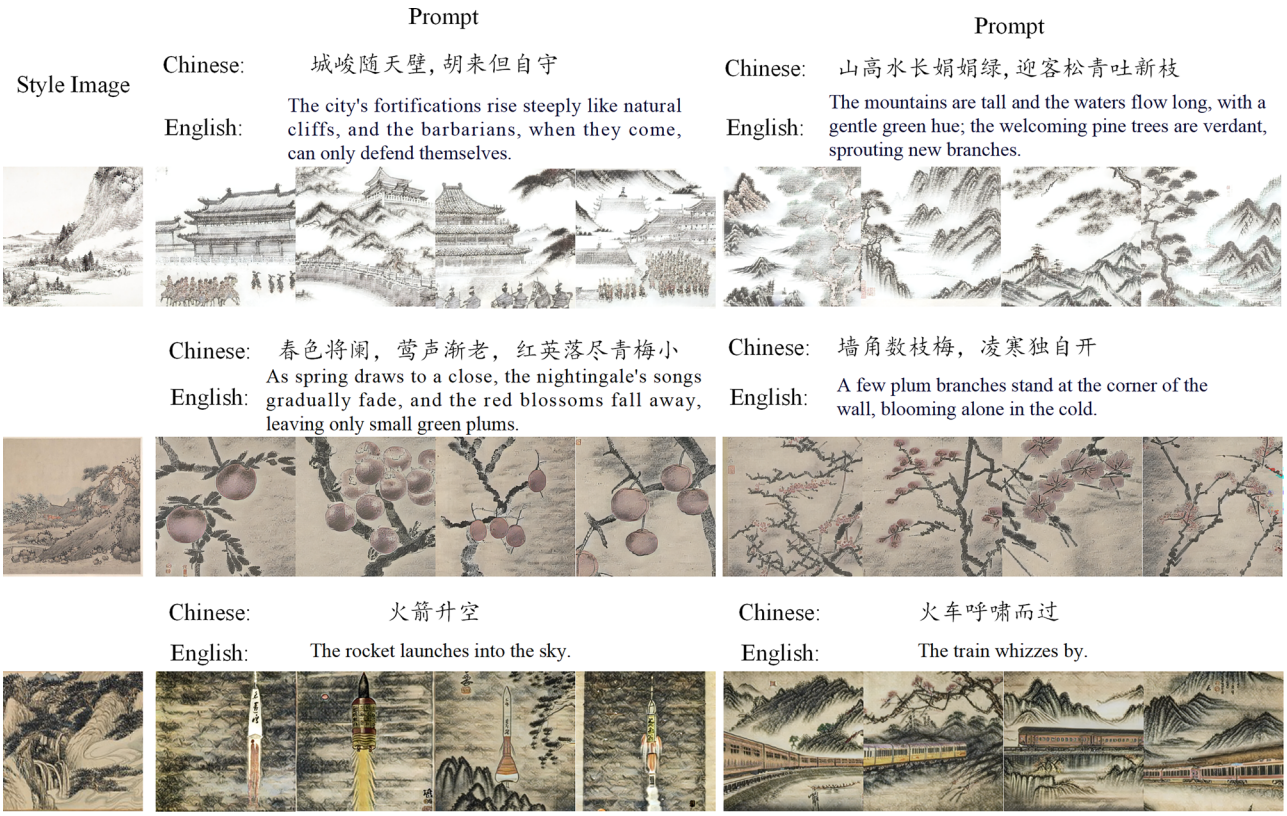

Fig. 5 | The results of Imagest on text and Chinese ink style conditioned art painting generation. Each row (left to right): style reference image, bilingual (Chinese/English) text prompt, and the generated pure ink painting, demonstrating strong comprehension of poetic imagery and authentic ink-wash rendering.

Fig. 6 | The Comparison results of Imagest and baseline methods on text-only conditioned art painting generation and synthesis. Columns from left to right: style reference (Style), text description (Description), our result (Ours), and results of baseline methods (Stable Diffusion, DALL·E 2, Imagen) generated directly from the same prompt.

creating novel art paintings using text that incorporates stylistic descriptions is a challenging task, regardless of the text's position within the sentence.

Furthermore, we used five prevalent fine-tuning methods -namely DreamBooth[31], Textual Inversion[44], InST[45], IP-Adapter[46], and LoRA[47] to fine-tune our base model, Stable Diffusion 1–4[15]. Specific style images were paired with the corresponding keywords and fed into the fine-tuning network for training.

Since different fine-tuning methods optimize different parts of the model (e.g., Textual Inversion only trains embedding vectors and thus converges faster, while DreamBooth performs full-model fine-tuning and

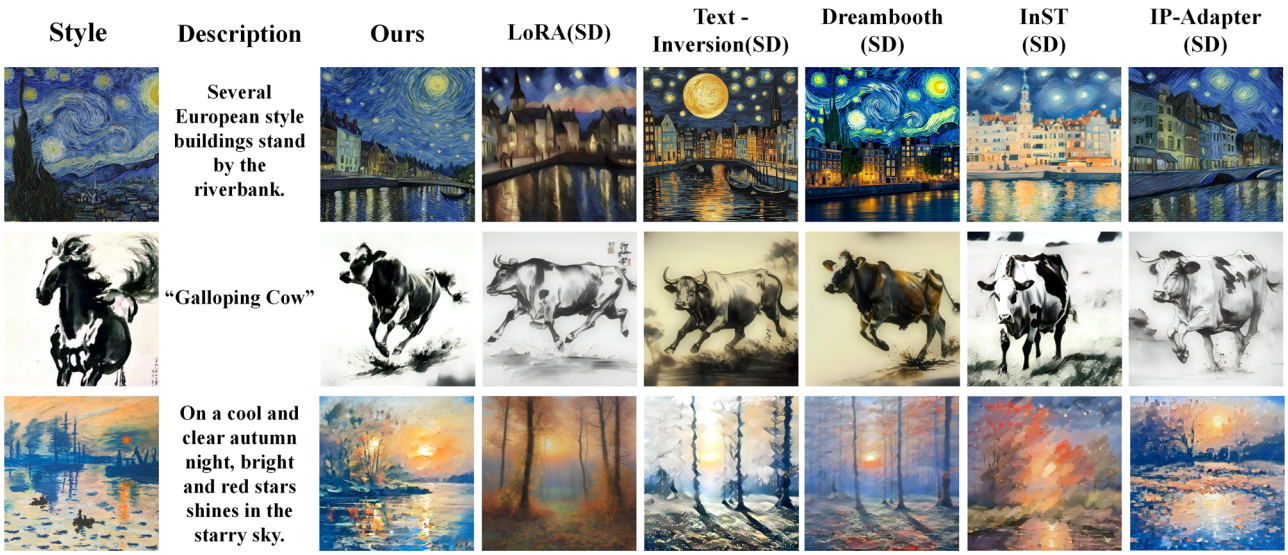

Fig. 7 | The Comparison results of Imagest and fine tune baseline methods on art painting generation and synthesis. Columns from left to right: style reference (Style), text description (Description), our result (Ours), and results of fine-tuned Stable Diffusion 1.5 methods (LoRA(SD), Text-Inversion(SD), Dreambooth(SD), InST(SD), IP-Adapter(SD)).

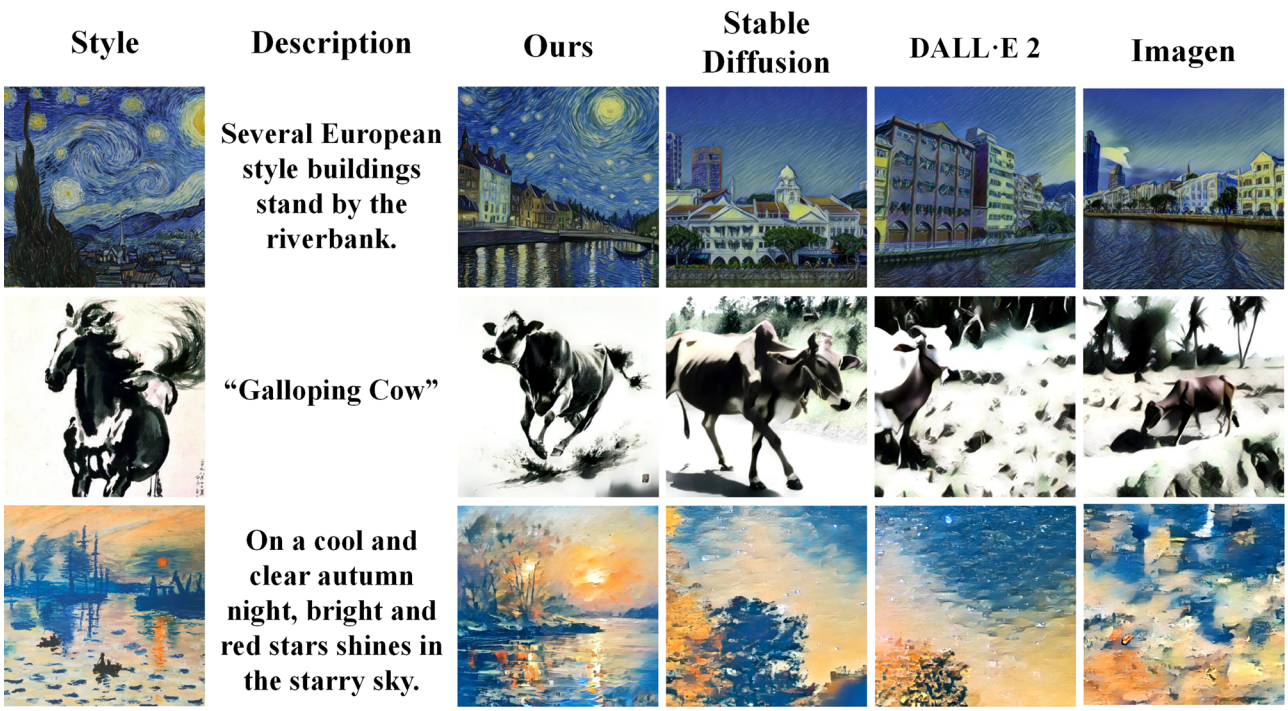

Fig. 8 | The Comparison results of Imagest and baseline methods with style transferring on art painting generation and synthesis. Columns from left to right: style reference (Style), text description (Description), our result (Ours), and results of Stable Diffusion, DALL·E 2, and Imagen after applying post-hoc style transfer networks.

is more time-consuming), we used the style images as training data to fine-tune SD 1.4 with all methods. To avoid overfitting, training was stopped as soon as each method reached convergence: Textual Inversion[44] converged around 3600 steps, LoRA[47] around 8400 steps, DreamBooth[48] around 9300 steps, InST[45] around 3600 steps and IP-adapter[46] around 4000 steps.

Following the training process, we used these fine-tuned baseline methods to synthesize art paintings for comparative analysis. A comparison of the experimental results with the fine-tuning methods is shown in Fig. 7. We also employed the three baseline models independently for the initial stage of text-only conditional image generation, producing preliminary results. Subsequently, these results were fed into the StyTR^2 framework[4] for style transfer, enabling us to obtain comparable results. The experimental results are presented in Fig. 8. The relevant parameter settings for SD 1.4 are provided in the "Data availability" section, where the location of the configuration file is specified Fig. 9.

As shown in the figures, in all artistic styles, the images generated by our Imagest method consistently demonstrate stylistic characteristics that closely resemble the original artworks. In contrast, all baseline models that rely solely on text prompts[11,12,15,24,25] fail to synthesize convincing styles, regardless of the input description.

Among the fine-tuning-based approaches[44–48], the recently proposed IP-Adapter shows strong competitive performance and achieves promising results. However, it still exhibits slight deviations in color fidelity and fine brushstroke details when compared to our method. Other fine-tuning

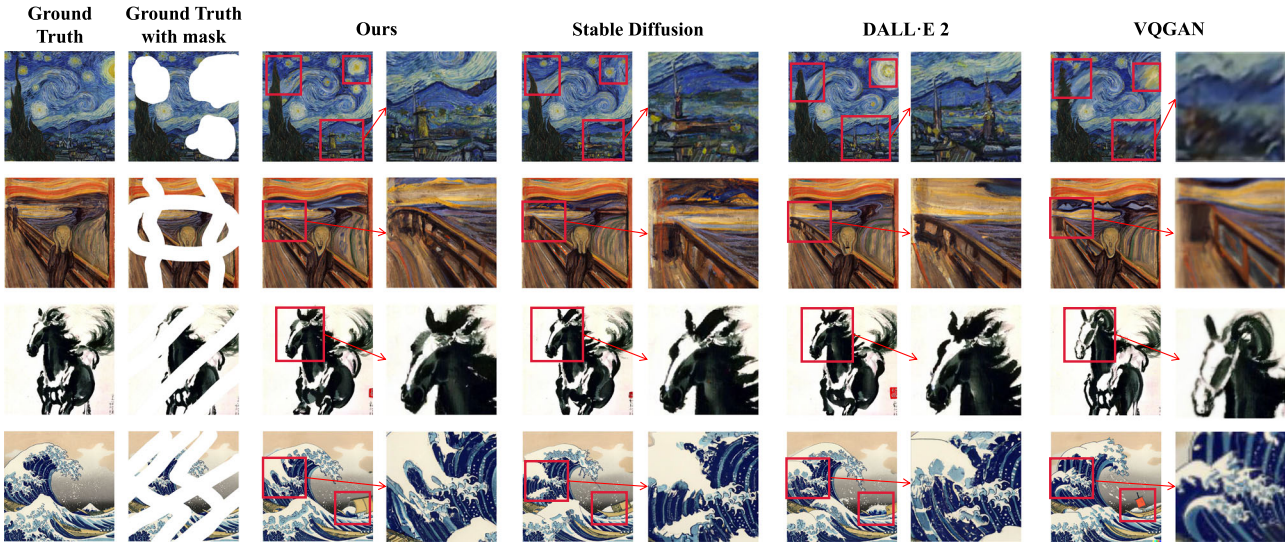

Fig. 9 | Selected results comparing Imagest with several baseline methods in the art paintings restoration task. For each method, two columns are shown: left is the full repaired image, right is the zoomed-in masked region (red box).

Table 1 | Comparison of Imagest and baselines on Style loss, CLIP score, FID, single-image, single-pass inference time (s), and peak VRAM (GB)

Model	Imagest (Ours)	Baselines without fine-tune			Baselines without fine-tune with Style Transfer		
		Stable Diffusion	DALL.E 2	Imagen	StableDiffusion	DALL.E 2	Imagen
Style loss ↓	**1.43**	5.89	4.97	5.43	1.74	1.79	1.72
CLIP score ↑	**78.36**	54.75	69.25	73.26	43.72	62.11	69.58
FID ↓	17.03	18.62	16.72	**14.99**	28.62	21.64	19.82
Inference time (s) ↓	3.63	**3.52**	11.23	16.69	3.71	12.08	17.83
Peak VRAM (GB) ↓	3.89	**3.25**	—	—	4.73	—	—

Model	Imagest(Ours)	Baselines with fine-tune without Style Transfer				
		Dreambooth (SD)	LoRA (SD)	InST (SD)	Textual Inversion (SD)	IP-Adapter (SD)
Style loss ↓	**1.43**	1.85	2.51	2.01	2.72	1.68
CLIP score ↑	**78.36**	67.83	65.62	69.42	62.01	76.46
FID ↓	17.03	19.52	19.87	18.32	20.46	**14.72**
Inference time (s) ↓	3.63	**3.57**	3.68	6.41	3.72	3.86
Peak VRAM (GB) ↓	3.89	3.52	**3.47**	4.25	3.62	4.06

Bold values indicate the best performance for each metric.

methods suffer from localized optimization, often capturing only partial stylistic features and failing to maintain global consistency.

Finally, in the two-stage generation methods[4] that combine initial content generation with style transfer, we observed that the style transfer process tends to distort the high-quality content image. The blending of content and style often leads to degradation in image quality, even though the overall color tone and stroke texture appear comparable to those of our approach.

Based on the experimental results, we can draw the conclusion that the art paintings synthesized using our method not only adhere closely to the descriptions provided by textual cues, but also possess a unique and distinct style. In contrast, other methods often fall short in either the precision of the generated content or the similarity of the art style to the original picture. Our approach thus demonstrates a superior balance between accuracy in interpretation of textual cues and creativity in artistic style.

Ultimately, we conducted a quantitative comparison of our method against several baseline approaches. In terms of evaluating the style of the synthesized art paintings, we employed style loss as a quantitative metric. Specifically, style loss is computed as the mean square error between the Gram matrices of the feature maps derived from the generated image and the target style image within the same layer of a pre-trained network[3]. This metric primarily captures the structural relationships among image features, such as lines and color blocks. For a quantitative assessment of content similarity, we utilized the CLIP-score, which serves as a benchmark for evaluating the correspondence between text and images. The CLIP-score is obtained by feeding text and image pairs into the CLIP model[16], converting them into feature vectors, and then computing the cosine similarity between these vectors. Additionally, we evaluated the quality of the generated images using the Fréchet Inception Distance (FID), which quantifies the quality and diversity of the images by comparing the distributional discrepancies between the generated images and real images within a predefined feature space. The experimental results are shown in Table 1.

Under the constraint of neither fine-tuning nor style transfer, the generation of compliant images solely relying on textual descriptions, which may be stylistically disparate or even entirely unrelated, while ensuring content alignment poses a significant challenge[11,12,15]. In the realm of fine-tuning-based approaches, the incorporation of styles such as “Starry Moon and Night” within the Dreambooth[48] and LoRA[47] frameworks primarily

localizes the style within the image, manifesting itself as a photographic appearance rather than exhibiting the brushstroke characteristics of an oil painting. Furthermore, these three methodologies exhibit dissimilarity to the style image "Starry Moon Night" in terms of color characteristics[44,47,48]. A similar issue arises in the coloration of results obtained through the Text Inversion[44] method in ink paintings and LoRA's[47] outcomes in the "Sunset Impression" style. The underlying cause of these results stems from the inherent limitation of prior fine-tuning methods, which are typically tailored for localized image editing, lacking the capability to recognize and adjust the image holistically. Both IP-Adapter[46] and InST[45] demonstrated strong competitiveness in terms of quantitative metrics, comparable to our method. However, they exhibited minor shortcomings in certain details, such as subtle deviations in color fidelity from the reference style images, which resulted in higher style loss scores. Nevertheless, both methods were still able to perform the tasks effectively. Conversely, our proposed Imagest demonstrates superior performance across various styles and exhibits closer proximity to the source image in terms of brushstrokes, colors, and other stylistic attributes compared to the baseline method. This is attributed to the integration of style features extracted by a pre-trained Swin Transformer during the noise prediction phase, where Swin's architecture inherently facilitates the integration of both global and local information[14]. Upon comparing the results with the two-stage generation process[4,11,12,15], we observe that the baseline method exhibits a lack of coordination in amalgamating content with style, rigidly overlaying the style onto the content without seamless integration. It is important to note that style transcends mere stylistic attributes, such as color and line, constituting a complex interplay with the depicted objects. In contrast, our methodology demonstrates enhanced coordination and superior image quality. This disparity arises because the two stages of image generation in the baseline method-content creation and style application-operate independently, with content being generated first and style subsequently imposed as a superficial overlay. In contrast, our approach incorporates style during the noise prediction phase, thereby integrating style fusion concurrently with content generation. This results in a more harmonious amalgamation of style and content, yielding results that are stylistically cohesive and authentic[15]. The values in the quantitative analysis also prove this point.

For runtime and memory analysis (single-image, single-pass), under a unified setting (batch size 1, 256 × 256, 30 DDIM steps, single RTX 4090D GPU), the measured inference latency and peak GPU VRAM of Stable Diffusion-based pipelines (plain SD[15], DreamBooth[31], LoRA[47], Textual Inversion[45], IP-Adapter[46]) fall within a narrow band: latency differences are on the order of a few tens of seconds, and memory footprints differ by only a few hundred megabytes-practically negligible for deployment decisions. By contrast, the two large hosted models (DALL.E 2[11] and Imagen[12]) exhibit seconds-level end-to-end latency and non-comparable VRAM (cloud-side), reflecting their substantially larger cascaded architectures. Within the SD family, our method maintains this lightweight profile while delivering lower latency and/or peak VRAM than most fine-tuning-based baselines, offering a balanced efficiency-quality trade-off that is well suited to conservation workflows with single-GPU constraints and batch processing needs.

The experimental results demonstrate that our proposed Imagest method effectively synthesizes images in a style that closely approximates the target style image, even with limited computational resources. Furthermore, the content of the generated images exhibits a high degree of alignment with the textual cue descriptions. In terms of image quality, our method ranks only second to DALL.E 2 and Imagen, which have been extensively trained on hundreds of A100 GPUs over extended periods spanning hundreds of days.

Style-guided painting restoration

Art paintings restoration is fundamentally distinct from image generation within the realms of art and computer vision. By substituting the text prompt constraint of Imagest with a masked image, where the mask denotes the damaged portions of the artwork, Imagest demonstrates proficiency in executing the art painting restoration task. Analogous to the LDM, for the restoration task, we refrain from utilizing the cross-attention mechanism across the entire latent space. Instead, we introduce style features derived from the style encoder model solely within the initial U-net layer[36,37], thereby enabling Imagest to accomplish the art paintings restoration task.

For binary mask generation over irregular regions, we compare the clean image with its manually damaged version to obtain a per-pixel difference map, apply Otsu thresholding to produce an initial binary result, refine it using morphological opening and closing to denoise and connect fragments, remove tiny components, and lightly dilate by 1–3 pixels to cover uncertain boundaries. This yields a binary free-form mask aligned with the input resolution, where 1 denotes damaged areas and 0 denotes intact areas. We perform latent-space inpainting by updating only masked latents with per-step copy-back for unmasked regions, followed by decoding and light edge blending, which yields context-consistent restorations confined to the hole.

Beyond numerical evaluation, we also conducted qualitative comparisons to visually assess the fidelity and consistency of restoration results.

To compare the performance of Imagest against other baseline models[11,15,49] in art paintings restoration, we curated a dataset of 100 renowned artworks, centrally cropped or resized to 256 × 256 pixels. These artworks were subsequently masked in varying locations and degrees and processed through each model for restoration. The baseline models included diffusion models that have exhibited strong performance in both image editing and generation, such as DALL.E 2[25], Stable Diffusion[15], and VQGAN[49]. A selection of results is depicted in Fig. 6. Visually, the diffusion models outperformed VQGAN. DALL.E 2[11] also delivered commendable restoration results, although with training challenges. Stable Diffusion[15] overcame training difficulties, but lacked precision and creativity in finer details. VQGAN[49] suffered mode collapse during large-scale training and produced images of inferior quality compared to diffusion models. Our Imagest not only excelled in restoring art paintings with specific stylistic constraints but also did so without requiring extensive resources, presenting vast potential for application. The results further demonstrated Imagest's superiority in restoring heavily damaged artworks, producing outputs that were closer to the Ground Truth in both detail and imagination.

In addition, we compared the restored detail regions of each damaged image with different methods. Our Imagest method and DALL.E 2[11] both demonstrate strong restoration capabilities, recovering both the stylistic features and semantic content of the original artworks. Stable Diffusion[15] shows slightly weaker performance in terms of content accuracy, occasionally missing finer structural elements. VQGAN[49], meanwhile, tends to produce blurrier results, and the overall generated images differ more noticeably from the original in both appearance and style.

To further substantiate the robustness of our model, we restored a set of artworks spanning varied damage extents and spatial coverage and compared performance against four representative baselines. Specifically, we report the average FID[50,51] and perceptual loss LPIPS[52], complemented by SSIM[53] and PSNR (dB)[54] computed within the restored regions (higher is better for SSIM/PSNR; lower is better for FID/LPIPS). As summarized in Table 2, DALL.E attains the best scores on FID, LPIPS, and SSIM, yet it is not the top performer on PSNR, a likely consequence of its outputs exhibiting slight over-smoothing, which can depress pixel-level fidelity despite favor-

Table 2 | The FID, LPIPS, SSIM, and PSNR of Imagest and several baseline methods

Model	Imagest	VQGAN	Stable Diffusion 1–4	Stable Diffusion 1–5	DALL.E 2
FID↓	<u>10.62</u>	12.89	12.30	10.87	**10.03**
LPIPS↓	<u>0.244</u>	0.280	0.252	0.249	**0.225**
SSIM ↑	<u>0.882</u>	0.835	0.864	0.879	**0.892**
PSNR ↑	**27.5**	24.3	25.8	26.9	<u>27.1</u>

Bold values indicate the best performance under each metric, while underlined values denote the second-best.

able perceptual metrics. Our method ranks second overall across the metrics, offering a balanced trade-off between restoration quality and computational cost, which is particularly pertinent for conservation workflows operating under single-GPU constraints and batch processing needs.

Discussion

This study set out to explore the potential of Imagest, a multiconditioned LDM, in the tasks of style-guided image generation and restoration of artwork. The experimental findings, illustrated in Figs. 4–8 and detailed in Tables 1–2, suggest that Imagest offers competitive performance on several evaluation criteria. In what follows, we discuss the results in light of current baselines, examine contributing factors, and outline limitations and possible future directions.

Firstly, we will conduct an analysis of the experimental results.In evaluating style preservation, it is observed that non-fine-tuned baseline models [11,12,15], which rely solely on textual input for stylistic control, often struggle to establish an accurate correspondence between descriptive cues and stylistic rendering. This may be attributed to the inherent ambiguity in mapping textual semantics to visual style, particularly in the absence of explicit style image inputs. In contrast, approaches based on fine-tuning[31,44,47], while incorporating reference style images during training, tend to focus on subject-centric adaptation. As a result, the learned style is frequently confined to localized regions rather than being globally distributed across the image. Moreover, methods that operate primarily within the latent space are susceptible to feature compression, which can diminish the fidelity of stylistic representation, especially in complex or highly textured scenarios.

With respect to content-style alignment, both fine-tuned and non-fine-tuned baselines occasionally exhibit a degree of semantic inconsistency. In particular, the fusion between style cues and object representations can be suboptimal, leading to outputs in which key elements are misaligned or visually ambiguous. The approach proposed in this work seeks to mitigate such issues by leveraging a Swin Transformer-based style encoder, which captures multiscale visual features while preserving structural coherence. The use of hierarchical attention further facilitates the integration of global context and local detail. Additionally, the model undergoes an extended fine-tuning phase, allowing for more stable adaptation between textual semantics and visual style. This may contribute to the relatively coherent fusion observed in the generated outputs.

Regarding overall image quality, both Imagest and several baseline methods[15,31,44,47] operate within a latent diffusion framework. While this paradigm offers computational efficiency by avoiding operations in high-dimensional pixel space, it inevitably introduces a degree of compression-related information loss[11,12]. Nonetheless, the experimental results suggest that, when guided by structured feature injection and appropriately conditioned attention, latent space generation can still produce visually plausible results within a constrained computational budget. However, further work is needed to better understand the trade-off between efficiency and fidelity, particularly in high-resolution or style-intensive generation tasks.

In the context of the image restoration task, Imagest was further assessed for its ability to reconstruct damaged paintings while preserving stylistic integrity. In addition to its generative capabilities, Imagest was also evaluated in the context of style-guided painting restoration, a task that differs fundamentally from image generation in its objectives and constraints. For this task, the text prompt on the input was replaced with a partially masked image representing the damaged artwork, while style guidance was retained through a reference image. The restoration mechanism does not apply cross-attention across the entire latent space; instead, the style features extracted by the Swin encoder are introduced selectively within the initial layer of the U-Net architecture[36,37], allowing the model to incorporate stylistic priors early in the denoising process.

Qualitative comparisons (Fig. 6) suggest that diffusion-based models generally outperform VQGAN in visual fidelity, with Imagest producing more stylistically consistent and structurally coherent restorations, particularly in cases involving severe degradation. DALL.E 2 also demonstrated notable performance in the restoration task. Stable Diffusion restored structural elements effectively but showed limitations in fine-grained artistic expression. Imagest achieved comparable or improved visual outcomes while operating under more constrained resources.

Quantitative evaluation using these metrics (Table 2) further supports these observations. Imagest exhibited lower FID and perceptual loss compared to all baselines tested, suggesting that it maintains both statistical similarity and perceptual quality in restored outputs. These results indicate that the proposed framework is not only adaptable to synthesis tasks but also holds promise for practical restoration scenarios where both visual coherence and style preservation are essential.

It should be noted that our observations align with the broader trend: larger foundation models-such as DALL.E 2 and Imagen-tend to deliver higher perceptual quality across both image generation and restoration tasks[11,12]. However, these gains typically presuppose substantial compute budgets (multi-GPU/TPU clusters) and high training/inference costs, which remain impractical for individual researchers or small conservation labs. In contrast, the Stable Diffusion ecosystem offers an accessible path for single-GPU users, but generic checkpoints are not ideal for personalized restoration without nontrivial adaptation. Our work therefore focuses on a single, well-defined restoration task, aiming for a balanced trade-off: we accept a small quality gap relative to ultra-large models in exchange for dramatically lower computational and deployment cost, prioritizing reproducibility and practicality over absolute compute dominance.

While Imagest demonstrates promising performance across multiple tasks, several practical challenges remain. First, the model's reliance on a 24 GB GPU during training may limit accessibility for users with more constrained computational resources, posing a barrier to broader adoption in academic or resource-limited settings. Second, the average inference time of ~3–5 s per image complicates deployment in interactive or real-time applications; lowering latency would better support iterative user interaction and high-throughput generation. Third, under complex or lengthy textual prompts, the model can exhibit reduced semantic coherence-manifesting as inaccurate object counts, misrepresented spatial relationships, or loss of fine-grained details-suggesting the need for stronger prompt understanding and conditioning mechanisms. Fourth, we did not include experiments on authentically damaged collection items requiring conservation treatment, because we currently lack a sufficiently large, rights-cleared dataset for training and validation; assembling such a corpus with curator-verified masks and standardized imaging is non-trivial and will be a priority in our follow-up work.

In museum deployment and other industrial production scenarios, although diffusion-based image generation and restoration methods excel in stylistic fidelity and visual quality, their practical implementation remains limited by the dependence on high VRAM hardware. The current design of Imagest, leveraging Stable Diffusion and lightweight fine-tuning strategies, already represents a pragmatic compromise between performance and resource demand. To further enable deployment in museum restoration, virtual exhibition, or creative production pipelines, we envision several potential directions for optimization: (1) employing model compression techniques (e.g., pruning or knowledge distillation) to reduce memory consumption; (2) converting the model into efficient inference formats such as ONNX or TensorRT to support edge-device compatibility; and (3) implementing batch generation or asynchronous inference scheduling to improve throughput and latency. These improvements will help meet the efficiency and deployment constraints faced in real-world cultural heritage and industrial applications.

Looking ahead, Imagest holds promise for both the conservation of digital heritage and the design of contemporary art. Its ability to synthesize and restore stylistically coherent images from minimal input makes it suitable for tasks such as reconstructing damaged artifacts, generating style-consistent virtual exhibitions, and producing educational materials that preserve artistic context. At the same time, its flexibility supports creative workflows in art and design, allowing artists and designers to rapidly prototype visual concepts, explore cross-style

recomposition, and produce original content for media and commercial applications. In summary, we introduce a novel model, termed Imagest, tailored for the task of synthesizing and restoring art paintings conditionally on style and textual prompts. This model encompasses two primary modules: the LDM module and the feature cueing module. Thanks to its sophisticated style extractor and robust diffusion model, Imagest demonstrates exceptional performance in the realm of image generation and stylization tasks. Furthermore, our extensive evaluations of the WikiArt dataset and a Chinese landscape painting dataset reveal that Imagest achieves impressive results in restoration and composition tasks while utilizing comparatively modest computational resources.

Data availability

All artistic data used in this study were sourced from three publicly available datasets: WikiArt, ChipPhi, and TCLPD. The WikiArt dataset is a widely adopted benchmark for artistic image generation and style transfer, comprising over 80,000 artworks across 27 distinct styles and numerous artists, covering a broad spectrum of historical periods and artistic movements. The ChipPhi and TCLPD datasets focus on traditional Chinese ink painting, featuring curated collections that emphasize brushwork, symbolic composition, and esthetic structure. All fine-tuning samples were uniformly resized to a resolution of 256×256 pixels to maintain consistency across training. Finally, we randomly split every dataset into 85\% for training and 15\% for testing. Here are the repositories of these datasets where you can download them: WikiArt dataset: A widely used benchmark containing over 80{,}000 artworks across 27 artistic styles and numerous historical periods. https://www.wikiart.org/. TCLPD (Traditional Chinese Landscape Painting Dataset): Proposed in End-to-End Chinese Landscape Painting Creation Using Generative Adversarial Networks https://github.com/alicex2020/Chinese-Landscape-Painting-Dataset. TCLPD (Traditional Chinese Landscape Painting Dataset): From End-to-End Chinese Landscape Painting Creation Using GANs.https://github.com/alicex2020/Chinese-Landscape-Painting-Dataset. CHIPPHI dataset: A curated dataset of Chinese ink paintings. Download link: https://pan.baidu.com/s/1oXFVv1tZCkUSoH2pSxWFSA (password: nqhi), Project page: https://github.com/PKU-IMRE/ChipGAN. The backbone model used in our framework is Stable Diffusion v1.4, a latent diffusion model pretrained on the large-scale LAION-2B dataset containing image-text pairs. The pretrained weights are publicly released by the CompVis research group and are accessible via the Hugging Face platform: https://huggingface.co/CompVis/stable-diffusion-v1-4. The model can be loaded using the diffusers library or executed using scripts from the official GitHub repository. Model architecture parameters are detailed in the v1-inference. yaml configuration file available under the configs directory. For text encoding, we adopt CLIP (Contrastive Language–Image Pretraining), a vision-language model developed by OpenAI. CLIP was trained on 400 million image–text pairs using contrastive learning to embed visual and textual information into a shared latent space. Pretrained CLIP models are available from OpenAI's official GitHub repository: https://github.com/openai/CLIP, and are also integrated into frameworks such as Hugging Face Transformers and OpenCLIP. The visual style encoder used in this study is based on the Swin-B (Base) architecture. It consists of four hierarchical stages with embedding dimensions of 96, 192, 384, and 768, respectively. The number of Swin Transformer blocks in each stage is set to 2, 2, 18, and 2. The input image resolution is 256×256, with a fixed window size of 7×7. Relative positional bias is applied within local attention windows, and patch merging is employed between stages for spatial downsampling.

Received: 13 January 2025; Accepted: 7 November 2025;
Published online: 28 November 2025

References

1. Elharrouss, O., Almaadeed, N., Al-Maadeed, S. & Akbari, Y. Image inpainting: a review. *Neural Process. Lett.* **51**, 2007–2028 (2020).
2. 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 Computer Society, Venice, Italy, 2017).
3. 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, Las Vegas, NV, USA, 2016).
4. Deng, Y. et al. StyTr^2: Image style transfer with transformers. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11326–11336 (IEEE, New Orleans, LA, USA, 2022). https://doi.org/10.1109/CVPR52688.2022.01105.
5. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T. & Efros, A. A. Context encoders: Feature learning by inpainting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 2536–2544 (IEEE, Las Vegas, NV, USA, 2016).
6. Iizuka, S., Simo-Serra, E. & Ishikawa, H. Globally and locally consistent image completion. *ACM Trans. Graph.* **36**, 1–14 (2017).
7. Goodfellow, I. et al. Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).
8. Reed, S. et al. Generative adversarial text to image synthesis. In *International Conference on Machine Learning* 1060–1069 (PMLR, New York, NY, USA, 2016).
9. Zhang, H. et al. Stackgan++: realistic image synthesis with stacked generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 1947–1962 (2018).
10. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).
11. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint* https://doi.org/10.48550/arXiv.2204.06125 (2022).
12. Saharia, C. et al. Photorealistic text-to-image diffusion models with deep language understanding. *Adv. Neural Inf. Process. Syst.* **35**, 36479–36494 (2022).
13. Kachkine, A. Physical restoration of a painting with a digitally constructed mask. *Nature* **642**, 343–350 (2025).
14. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE/CVF International Conference on Computer Vision* 10012–10022 (IEEE, Montreal, QC, Canada (virtual), 2021).
15. 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, New Orleans, LA, USA, 2022).
16. Radford, A. et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* 8748–8763 (PMLR, Virtual (hosted by Vienna, Austria), 2021).
17. Dawson, T. Examination, conservation and restoration of painted art. *Color. Technol.* **123**, 281–292 (2007).
18. Kubik, M. E. Preserving the painted image: the art and science of conservation. Journal of the International Colour Association 5, 1–8 (2010).
19. Heitzinger, T., Woedlinger, M. & Stork, D. G. Artist-specific style transfer for semantic segmentation of paintings: the value of large corpora of surrogate artworks. *Electron. Imaging* **34**, 1–6 (2022).
20. Stork, D. G. *Pixels & Paintings: Foundations of Computer-Assisted Connoisseurship* (John Wiley & Sons, Hoboken, NJ, USA, 2023).
21. Elad, M. & Milanfar, P. Style transfer via texture synthesis. *IEEE Trans. Image Process.* **26**, 2338–2351 (2017).
22. Salimans, T. et al. Improved techniques for training GANs. *Adv. neural Inf. Process. Syst.* **29**, 2234–2242 (2016).
23. Zhang, H. et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proc. IEEE International Conference on Computer Vision* 5907–5915 (IEEE, Venice, Italy, 2017).

24. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. In 2nd International Conference on Learning Representations (ICLR 2014).
25. Ramesh, A. et al. Zero-shot text-to-image generation. In *International Conference on Machine Learning* 8821–8831 (PMLR, Virtual (hosted by Vienna, Austria), 2021).
26. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training (OpenAI, 2018).
27. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 9 (2019).
28. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
29. Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. In *9th International Conference on Learning Representations* (ICLR 2021, Virtual Event, Austria, 3–7 May 2021).
30. Kawar, B. et al. Imagic: text-based real image editing with diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 6007–6017 (IEEE, Vancouver, BC, Canada, 2023).
31. Ruiz, N. et al. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 22500–22510 (IEEE, Vancouver, BC, Canada, 2023).
32. Nichol, A. et al. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In Proceedings of the 39th International Conference on Machine Learning, 16784–16804 (PMLR, Baltimore, MD, USA, 2022).
33. Vaswani, A. et al. Attention is all you need. In Advances in Neural Information Processing Systems 30, 5998–6008 (Curran Associates, Inc., Long Beach, CA, USA, 2017).
34. Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. In 9th International Conference on Learning Representations (ICLR 2021) (OpenReview.net, Virtual Event, Austria, 2021).
35. Dosovitskiy, A. & Brox, T. Generating images with perceptual similarity metrics based on deep networks. *Advances in Neural Information Processing Systems* **29** https://doi.org/10.48550/arXiv.1602.02644 (2016).
36. Cao, H. et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, 205–218 (Springer, 2023).
37. Fan, C.-M., Liu, T.-J. & Liu, K.-H. Sunet: swin transformer unet for image denoising. In *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, 2333–2337 (IEEE, 2022).
38. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, Las Vegas, NV, USA, 2016).
39. He, K., Zhang, X., Ren, S. & Sun, J. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 630–645 (Springer, 2016).
40. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 5485–5551 (2020).
41. Phillips, F. & Mackintosh, B. Wiki art gallery, inc.: a case for critical thinking. *Issues Account. Educ.* **26**, 593–608 (2011).
42. He, B., Gao, F., Ma, D., Shi, B. & Duan, L.-Y. Chipgan: a generative adversarial network for chinese ink wash painting style transfer. In *Proc. 26th ACM International Conference on Multimedia* 1172–1180 (ACM, Seoul, Korea, 2018).
43. Xue, A. End-to-end chinese landscape painting creation using generative adversarial networks. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision* 3863–3871 (IEEE, Virtual, 2021).
44. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G. & Cohen-Or, D. An image is worth one word: personalizing text-to-image generation using textual inversion. In 11th International Conference on Learning Representations (ICLR 2023) (OpenReview.net, Kigali, Rwanda, 2023).
45. Zhang, Y. et al. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10146–10156 (IEEE, Vancouver, BC, Canada, 2023).
46. Ye, H., Zhang, J., Liu, S., Han, X. & Yang, W. IP-Adapter: text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint* https://doi.org/10.48550/arXiv.2308.06721 (2023).
47. Hu, E. J. et al. Lora: low-rank adaptation of large language models. *arXiv preprint* https://doi.org/10.48550/arXiv.2106.09685 (2021).
48. Ruiz, N. et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 22500–22510 (IEEE, 2023).
49. Esser, P., Rombach, R. & Ommer, B. Taming transformers for high-resolution image synthesis. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 12873–12883 (IEEE, Nashville, TN, USA, 2021).
50. Kang, M. & Park, J. Contragan: contrastive learning for conditional image generation. *Adv. Neural Inf. Process. Syst.* **33**, 21357–21369 (2020).
51. 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. *Adv. Neural Inf. Process. Syst.* **30**, 6626–6637 (2017).
52. Suvorov, R. et al. Resolution-robust large mask inpainting with fourier convolutions. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision* 2149–2159 (IEEE, 2022).
53. 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).
54. Horé, A. & Ziou, D. Image quality metrics: PSNR vs. SSIM. In *Proc. 20th International Conference on Pattern Recognition (ICPR)* 2366–2369 (IEEE, Istanbul, Turkey, 2010).

Acknowledgements

Many thanks to the authors for their professional advice on revisions, which led to a significant improvement in the quality of the article. This research was funded by the National Natural Science Foundation of China (62271393 and 62571051), Xi'an Science and Technology Programme Demonstration Project on Science and Technology Innovation for Social Development (24SFSF0002), Archaeological Talent Promotion Program of China (2024-267), Natural Science Basic Research Program of Shaanxi (Program No.2025JC-QYXQ-039).

Author contributions

Conceptualization: J.H., Y.Z. Methodology: J.H., Y.Z., P.Z. Formal analysis: J.H., Y.Z., P.Z. Investigation: J.H., G.G. Resources: G.G. Data curation: Y.Z., G.G. Writing—original draft preparation: J.H., J.G., Y.Z., P.Z. Writing—review and editing: J.H., Y.Z., J.G. Supervision: J.H., Y.Z., P.Z., M.Z. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Yuhe Zhang or Guohua Geng.