Exploring Controllable Ability through Text-to-Image Diffusion Model for Painting-Style Design

Yifei Zhao , Ziqi Liang * , Yingrui Qiu , Xiaona Wang , Wanggong Yang *

Posted Date: 15 August 2024

doi: 10.20944/preprints202408.1035.v1

Keywords: diffusion model; painting-style design; text-to-image (T2I) models; computer-aided design

 doi:10.20944/preprints202408.1035.v1

Article

Exploring Controllable Ability through Text-to-Image Diffusion Model for Painting-style Design

Yifei Zhao, Ziqi Liang, Yingrui Qiu, Xiaona Wang and Wanggong Yang *

School of New Media, Beijing Institute of Graphic Communication, Beijing 102600, China; zhaoyifei@bigc.edu.cn
* Correspondence: ywg@bigc.edu.cn

Abstract: Painting style and creativity are fundamental to art, defining a work's uniqueness and expression. They embody an artist's personality, convey emotions, establish identity, and shape audience perception, adding depth and value to the work. Recently, diffusion models have gained popularity in art design, animation, and gaming, especially in original painting creation, poster design, and visual identity development. Traditional creative processes, while demanding unique imagination and creativity, often face challenges such as slow innovation, high reliance on manual efforts, high costs, and limitations in scalability. Consequently, exploring painting-style creative design through deep learning has emerged as a promising trend. In this paper, we introduce a novel network architecture, the Painting-Style Design Assistant Network (PDANet), for style transformation. To support this, we curated the Painting-42 dataset, comprising 4,055 works by 42 renowned Chinese painters. PDANet leverages this dataset to capture the aesthetic intricacies of Chinese painting, offering rich design references. Furthermore, we propose a lightweight Identity-Net for large-scale text-to-image (T2I) models, which aligns internal knowledge with external control signals, thereby enhancing the capabilities of existing T2I models. The trainable Identity-Net feeds image prompts into the U-Net encoder to generate diverse and stable images. Both quantitative and qualitative analyses demonstrate that our approach outperforms current methods, delivering high-quality generated content with wide-ranging applicability.

Keywords: diffusion model; painting-style design; text-to-image (T2I) models; computer-aided design

1. Introduction

Chinese painting is a traditional ancient art with a long history of more than 3,000 years, rich in content and diverse subject matter [1–6]. This ancient art form has diverse styles, divided into different dynasties and schools, reflecting the characteristics of traditional Chinese culture [7]. Compared to the labor-intensive and time-consuming nature of conventional hand-painting, computer-assisted image generation has led to a paradigm shift in efficiency and innovation [8]. By leveraging cutting-edge technology, artists can quickly produce complex effects, saving time and costs while expanding the boundaries of artistic expression and unlocking new creative possibilities [9]. Generative Adversarial Networks (GANs) represent the state-of-the-art in this field, revolutionizing the field of computer-assisted image generation [10]. However, despite their impressive capabilities, GANs often have difficulty maintaining the stability of the quality of generated images, which can appear blurry or distorted due to imperfect control over content generation [11]. In addition, GANs are susceptible to mode collapse, resulting in limited diversity in generated images, which can severely limit their applicability in real-world scenarios [12].

Recent breakthroughs in the field of generative art have opened up new avenues for artistic expression, including the use of large-scale models such as stable diffusion (SD) [13], Midjourney [14] and DALL-E 3 [15]. Among them, SD has shown remarkable strength in generating Chinese landscape paintings and creating high-quality works with complex details and coherent structures. In particular, the artworks generated by SD faithfully capture the texture and layering characteristics of traditional ink paintings, demonstrating its extraordinary ability to imitate the nuances of this ancient art form [16]. The success of SD can be attributed to its innovative approach, which combines style reference images with descriptive clues to provide precise control over style and content [17]. This enables artists to create works that meet specific artistic expectations while also allowing them to

 doi:10.20944/preprints202408.1035.v1

explore new creative possibilities [18]. In addition, SD's automatic generation process significantly shortens the creation time, allowing artists to focus on conceptualization and creative exploration [19]. SD uses digital technology to promote the seamless combination of traditional Chinese painting and modern technology, thus promoting innovation and protecting artistic heritage [20].

This paper introduces the Painting-42 dataset, a comprehensive collection of images and text annotations designed specifically for Chinese painting styles. Using this dataset, we introduce a groundbreaking model, the painting-style design assistance network (PDANet), which is characterized by its lightweight training parameters and its ability to guide text inputs to generate design references that meet predefined criteria. The PDANet model aims to address the challenges of text-to-image (T2I) generation, which aims to generate images from text prompts using big data and powerful computing power. While T2I generation has achieved high synthesis quality, it relies heavily on carefully crafted prompts and lacks flexible user control, which often results in an inaccurate reflection of the user's ideas. This limitation leads to inaccurate control and unstable results for non-expert users. To address this challenge, we propose Identity-Net, a lightweight model that excels at learning with relatively small datasets. Identity-Net aims to enhance the pre-trained T2I diffusion model by providing supplementary guidance. We hypothesize that a compact adapter model can effectively perform this role by mapping control information onto the internal knowledge of the T2I model instead of learning a completely new generation function. By doing so, Identity-Net enables more flexible and controllable T2I generation, allowing users to generate high-quality images that accurately reflect their ideas.

In summary, the primary contributions of this article are as follows.

- We present the Painting-42 dataset, a comprehensive collection of 4,055 works by 42 renowned Chinese painters from various historical periods. This high-quality dataset is specifically tailored to capture the intricacies of Chinese painting styles.
- We propose PDANet, a lightweight architecture for style transformation that pioneers the application of diffusion models in creative style design. PDANet excels in faithfully emulating the distinct styles of specific Chinese painters, enabling the rapid generation of painting-style designs directly from text inputs.
- We introduce the Identity-Net, a straightforward and efficient method that effectively adjusts the internal knowledge of T2I models and external control signals at minimal expense. Through comprehensive artistic evaluations conducted during user research, we demonstrate improved user preferences and validate the effectiveness of our approach.

2. Related Work

The advent of deep generative models has ushered in a new era of digital creation in the field of painting, with computer algorithms pioneering new approaches in the field of generative art [21]. However, controllable painting style generation remains a relatively underexplored area, hampered by the scarcity of relevant data and the limited adaptability of existing design frameworks [17]. To address this challenge, we propose PDANet, a new method for generating paintings with a specific style using a latent diffusion model [22]. By leveraging the power of deep learning, PDANet can produce high-quality, stylistically unique paintings that are both beautiful and controllable [23].

2.1. Generative Adversarial Networks

The convergence of artificial intelligence and artistic expression has inaugurated a new era of innovative applications in painting and art. Recent breakthroughs in computer-aided design have concentrated on developing sophisticated rendering and texture synthesis algorithms, particularly in the domain of image stylization. The advent of neural network-based style transfer methods, which synthesize images that amalgamate the style of one image with the content of another [24], has paved the way for novel creative possibilities. This paradigm has garnered considerable attention, sparking a further investigation into deep learning-based techniques, notably through the utilization of

 doi:10.20944/preprints202408.1035.v1

convolutional neural networks (CNNs) [25,26]. The efficacy of these approaches has been extensively demonstrated in various applications, showcasing their potential to revolutionize the field of artistic expression. Recently, researchers have investigated the incorporation of GANs into image style transfer [27–30], thereby expanding the frontiers of this field. GANs, a groundbreaking deep learning paradigm primarily employed for unsupervised learning and creative data generation, comprise two essential components: the generator G and the discriminator D [31]. Generator G learns to produce realistic novel data instances by mapping random noise to approximate the real-world data distribution, while discriminator D is trained to distinguish between authentic data and synthetic data generated by the generator. This interplay forms a dynamic game during training, driving both components to continually enhance their performance [32]. Ultimately, the objective is for the generator to produce high-fidelity samples that are indistinguishable from truthful data. Beyond image style transfer, GANs have found diverse applications in image restoration, music composition, video synthesis, and numerous other fields, underscoring their pivotal role in advancing generative model technology and pushing the boundaries of artificial intelligence.

The proposal of the sketch and paint GAN (SAPGAN) is a pioneering contribution in this research field [32], which plays an important role in generating sketches and paintings. This groundbreaking work introduced new concepts to the field, especially in the field of Chinese painting generation. Notably, SAPGAN represents the first end-to-end model designed to unconditionally create Chinese paintings, thus bridging the gap between artificial intelligence and art. The framework contains two basic components: a sketching-GAN that generates outlines and a paint-GAN that applies colors based on these outlines. By integrating these components, SAPGAN provides a powerful theoretical framework and practical methods for computer-generated artistic content. This technological breakthrough marks an important milestone in interdisciplinary research and highlights the potential of artificial intelligence (AI) to revolutionize artistic creation and push the boundaries of creative expression.

2.2. Painting Style Transfer

Style transfer entails the process of migrating stylistic elements from one image to another while preserving the content of the source image. This technique involves learning the intricate details of the target image and adapting them to the source image [33]. Recently, the application of generative artificial intelligence for text-guided image generation has gained significant traction, particularly with the advancements in large-scale diffusion models [34]. Notable examples include Glide [35], Cogview [36], Imagen [37], Make-a-scene [38], ediffi [39], and Raphael [40]. These approaches often employ self-distillation techniques in conjunction with adapters for guided generation, requiring minimal additional training while maintaining the fixed parameters of the original model. In the realm of image style transfer, GANs provide a robust framework. For instance, Pix2Pix was one of the earliest applications of GANs to image pair translation tasks, effectively achieving style transfer. Subsequent studies have further refined style transfer using GANs and their variants. CycleGAN, for example, handles unpaired data, while architectures such as Anycost GAN optimize resource utilization on diverse computing platforms. These advancements underscore the versatility and ongoing development of GAN-based techniques for optimizing style transfer effects.

Moreover, we explore an inversion-based style transfer method (InST) that harnesses the capabilities of diffusion models to conceptualize painting styles through a trainable textual representation. InST leverages inverse mapping techniques to accurately and efficiently extract and transfer artistic style characteristics from a source image. This approach enables the synthesis of novel stylized content without the need for elaborate textual guidance, operating effectively even with a single painting as a reference. By doing so, InST streamlines the style transfer process, allowing for greater flexibility and creative control.

 doi:10.20944/preprints202408.1035.v1

2.3. Diffusion Models

Diffusion models [41] have made significant progress in image synthesis by generating images from Gaussian noise via iterative denoising methods, which are based on rigorous physical principles governing the diffusion process and its reversal [42,43]. Recently, image diffusion models have attracted widespread attention in image generation [34,44]. Latent diffusion models (LDM) [13] perform the diffusion step within the latent image space [45], significantly reducing computational overhead. In the field of T2I generation, diffusion models leverage pre-trained language models such as CLIP [46] to encode textual inputs into latent vectors, resulting in state-of-the-art image synthesis results. It is worth noting that SD is an upscaled implementation of LDM, while Imagen [37] adopts a pyramid structure to directly diffuse pixels without involving the latent image, thus providing a unique approach to image synthesis.

Several recent approaches [39,47,48] have been proposed to modify cross-attention maps in pre-trained T2I models, enabling the steering of the generation process without requiring additional training. A notable advantage of these methods is their seamless integration with existing models. For instance, AFA [49] dynamically adjusts the contributions of multiple diffusion models based on various states, leveraging their strengths while suppressing their weaknesses. This approach differs from static parameter merging methods, offering a more adaptive and effective way to combine the capabilities of different models. In parallel efforts, Zhang et al. [50] have explored the use of task-specific control networks to facilitate conditional generation using pre-trained T2I models. This line of research has shown promise in enabling more precise control over the generation process. Furthermore, diffusion models have been successfully applied to a wide range of scenarios, including pose-guided person image synthesis [51], talking faces [52], virtual dressing [53], and story generation [54]. These applications demonstrate the versatility and potential of diffusion models in various domains. Therefore, it is essential to explore the potential of diffusion models in the synthesis of artistic style, a promising area of research that could benefit from the capabilities of these models.

3. Proposed Method

3.1. Preliminary

The T2I diffusion model, specifically SD [13], plays a crucial role in modeling the dynamic behavior and stable state distribution of complex systems by drawing an analogy between image generation and ink diffusion in water. This model comprises two primary processes, leveraging an auto-encoder and a modified UNet denoiser [55]. In the initial phase, SD employs an auto-encoder trained to encode images into a compact latent space, followed by reconstruction. This process enables the model to capture the essential features and patterns in the input data. The subsequent phase utilizes a modified UNet denoiser to directly refine this latent space, effectively removing noise and generating high-quality images. This streamlined approach can be formulated as follows:

$$\mathcal{L} = \mathbb{E}\mathbf{Z}t, \mathbf{C}, \mathbf{e}, t\Big(|\mathbf{e} - \mathbf{e}_\theta(\mathbf{Z}_t, \mathbf{C})|_2^2\Big), \tag{1}$$

where $\mathbf{Z}t = \sqrt{\overline{\alpha t}}\mathbf{Z}0 + \sqrt{1-\overline{\alpha t}}\mathbf{e}$ represents the noise-processed feature map in step t, with $\mathbf{e} \in \mathcal{N}(0, \mathbf{I})$ denoting the noise component. $\mathbf{C}$ represents the conditional information, and $\mathbf{e}_\theta$ is a function of the UNet denoiser, parameterized by θ.

During the inference process, the initial latent map $\mathbf{Z}_T$ is generated from a random Gaussian distribution. Given $\mathbf{Z}T$, the denoiser $\mathbf{e}\theta$ predicts noise estimates at each step t, conditioned on the conditional information $\mathbf{C}$. By subtracting these noise features, the noise map becomes increasingly refined, allowing for the recovery of the underlying signal. After t iterations, the resulting purified latent feature $\hat{\mathbf{Z}}_0$ is inputted into the decoder for image generation. In the conditional setup, SD leverages a pre-trained CLIP [46] text encoder to embed textual input into a sequence represented by $\mathbf{y}$. This embedded sequence is then incorporated into the denoising process using a cross-attention model, enabling the model to effectively integrate textual information into the image generation process.

 doi:10.20944/preprints202408.1035.v1

$$\begin{cases} \mathbf{Q} = \mathbf{W}_Q \phi(\mathbf{Z}_t),\ \mathbf{K} = \mathbf{W}_K \tau(\mathbf{y}),\ \mathbf{V} = \mathbf{W}_V \tau(\mathbf{y}), \\ \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right) \cdot \mathbf{V}. \end{cases} \tag{2}$$

Here, $\phi(\cdot)$ and $\tau(\cdot)$ are two trainable embeddings. $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are trainable projection matrices.

3.2. Identity-Net Design

We propose a multimodal diffusion model for the T2I architecture, in order to address the limitations of existing models in the generation of complex scenes. As illustrated in the upper part of Figure 1, relying solely on individual text inputs in PDANet can lead to instability and hinder accurate structural guidance for image synthesis. This issue is not due to the model's generation capabilities, but rather the challenge of providing precise generation guidance via textual inputs, which requires seamless alignment of internal knowledge from SD with external control. To overcome this limitation, we propose integrating multimodal inputs, which can be achieved at a low computational cost. This approach reduces the reliance on single conditions by enabling adapters to extract guidance features from various types of conditions. The pre-trained SD model keeps its parameters fixed, generating images based on input text features and additional guidance features. Using multimodal input, our model can effectively capture the complexities of the scene and provide more accurate structural guidance for image synthesis.

Our Identity-Net proposal, illustrated in the lower part of Figure 1, is designed to be simple, lightweight, and efficient. It comprises four feature extraction blocks and three downsampling blocks to adjust the feature resolution. Initially, the conditional input has a resolution of 512×512, which is downsampled to 64×64 using the pixel unshuffle operation [56]. At each scale, a convolutional layer and two residual blocks (RB) are employed to extract the conditional feature $\mathbf{F}c^k$. This process yields a multi-scale conditional feature $\mathbf{F}c = \mathbf{F}c^1, \mathbf{F}c^2, \mathbf{F}c^3, \mathbf{F}c^4$, where the dimensions match those of the intermediate feature $\mathbf{F}enc = \mathbf{F}enc^1, \mathbf{F}enc^2, \mathbf{F}enc^3, \mathbf{F}^4_{enc}$ in the encoder of the UNet denoiser. These features $\mathbf{F}c$ and $\mathbf{F}enc$ are subsequently combined at each scale using a combination operation. The process of conditional feature extraction and operations can be formulated as follows:

$$\mathbf{F}c = \mathcal{F}AD(\mathbf{C}), \tag{3}$$

$$\hat{\mathbf{F}}enc^i = \mathbf{F}enc^i + \mathbf{F}c^i,\ i \in 1, 2, 3, 4, \tag{4}$$

where $\mathbf{C}$ represents the conditional input, and $\mathcal{F}AD$ is the Identity-Net. Our proposed Identity-Net exhibits strong generalization capabilities and accommodates diverse structural controls, such as sketches, depth maps, semantic segmentation maps, and key poses. The condition maps corresponding to these modes are directly fed into task-specific adapters to extract condition features, denoted as $\mathbf{F}_c$.

In T2I tasks, CLIP plays a pivotal role in bridging the gap between textual descriptions and image representations. By leveraging both text encoders and image encoders, CLIP is trained using contrastive learning on a vast dataset of text-image pairs, thereby aligning patterns in the feature space. In the context of SD, employing a pre-trained CLIP text encoder to extract text embeddings from input text enables guidance of the denoising process. However, relying solely on text guidance may not sufficiently capture the intricate aesthetic characteristics inherent in Chinese painting. To mitigate this limitation, an additional image encoder is introduced to provide finer details, albeit necessitating parameter fine-tuning to adapt to painting-style design tasks. Our proposed PDANet method addresses these challenges by training adapters using frozen CLIP and SD, focusing on learning parameters specifically within linear layers, layer normalization, and cross-attention layers. This approach yields promising results without requiring extensive retraining, thereby streamlining the process.

 doi:10.20944/preprints202408.1035.v1

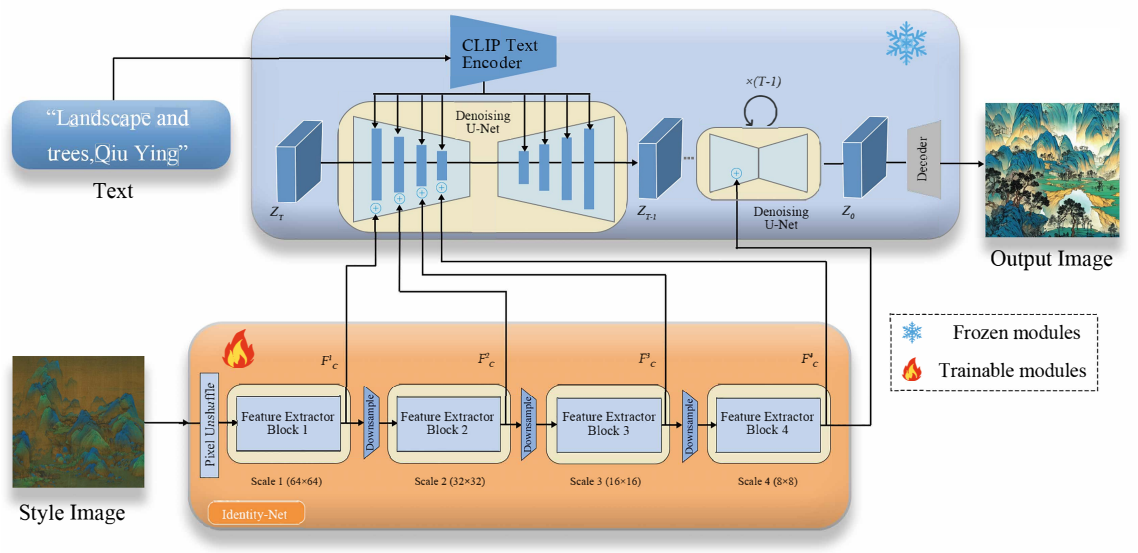

Figure 1. The overall framework of PDANet consists of two parts: (1) a pre-trained SD model with fixed parameters while freezing the parameters of other modules; (2) Identity-Net trains the internal knowledge and external control signals of the alignment model.

3.3. Model Optimization

During optimization, we maintain fixed parameters in SD while focusing on optimizing the Identity-Net. Each training sample consists of a triplet: the original image $\mathbf{X}_0$, the condition map $\mathbf{C}$, and the text prompt y. The optimization procedure mirrors that of SD. Specifically, starting with an image $\mathbf{X}_0$, we encode it into the latent space $\mathbf{Z}_0$ using the auto-encoder's encoder. Subsequently, we randomly select a time step t from the interval $[0, T]$ and introduce corresponding noise to $\mathbf{Z}_0$, yielding $\mathbf{Z}_t$. Mathematically, our Identity-Net is optimized as follows:

$$\mathcal{L}_{AD} = \mathbb{E}_{\mathbf{Z}_0,t,\mathbf{F}_c,\epsilon\sim\mathcal{N}(0,1)}\left[||\epsilon - \epsilon_\theta(\mathbf{Z}_t, t, \tau(\mathbf{y}), \mathbf{F}_c)||_2^2\right]. \tag{5}$$

Where $\mathbb{E}$ denotes the expectation over multiple samples. Here, $\mathbf{Z}_0$ represents the initial latent variable, t is the time step or iteration count, $\mathbf{F}_c$ signifies conditional features or contextual information, and $\epsilon \sim \mathcal{N}(0, 1)$ indicates that the noise ϵ is sampled from a standard normal distribution. The term $||\cdot||_2^2$ denotes the squared Euclidean norm, used to measure the error between the predicted noise ϵ_θ and the true noise ϵ. The model output $\epsilon_\theta(\mathbf{Z}_t, t, \tau(\mathbf{y}), \mathbf{F}_c)$ depends on the parameters θ, the latent variable $\mathbf{Z}_t$ at time t, the embedding $\tau(\mathbf{y})$ of the label $\mathbf{y}$, and the conditional features $\mathbf{F}_c$.

In the diffusion model, embedding time as a condition during sampling is crucial for effective guidance. Through our experiments, we have observed that incorporating time embedding into the adapter significantly enhances its guidance capabilities. However, this approach requires the adapter's involvement in every iteration, which contradicts our goal of simplicity and compactness. To address this issue, we employ strategic training methods to optimize the adapter's performance. Specifically, we segment the DDIM inference sampling process into three stages: early, middle, and late stages. We introduce guidance information at each of these stages to investigate its impact on the results. Interestingly, our findings reveal that adding guidance during the middle and late stages has a minimal impact on the results, suggesting that the primary content of the generated output is largely determined in the early sampling stage. Consequently, if t is sampled from the later stages, the guidance information tends to be ignored during training, leading to suboptimal performance. To bolster the adapter's training, we adopt a non-uniform sampling strategy to increase the likelihood of t falling within the early sampling stage. We utilize a cubic function, $t = (1 - (\frac{t}{T})^3) \times T,\ t \in U(0, T)$, where t is sampled from a uniform distribution. This distribution helps address the issue of weak guidance observed with uniform sampling, particularly in tasks such as color control. The cubic

 doi:10.20944/preprints202408.1035.v1

sampling strategy effectively mitigates these weaknesses, leading to improved overall performance of the diffusion model.

3.4. Inference Stage

During training, we exclusively optimize Identity-Net while maintaining the parameters of the pre-trained diffusion model fixed. The Identity-Net is trained on the dataset containing image-text pairs, adhering to the same training objective as the original SD:

$$L_{\text{simple}} = \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{\epsilon},\boldsymbol{c}_t,\boldsymbol{c}_i,t}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t,\boldsymbol{c}_t,\boldsymbol{c}_i,t\right)\|^2. \tag{6}$$

In addition, we randomly exclude image conditions during the training stage to facilitate guidance in the inference stage without relying on classifiers:

$$\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t,\boldsymbol{c}_t,\boldsymbol{c}_i,t) = w\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,\boldsymbol{c}_t,\boldsymbol{c}_i,t) + (1-w)\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t). \tag{7}$$

Here, we nullify the CLIP image embedding simply by setting it to zero when the image condition is excluded.

Since the text cross-attention and image cross-attention are separate, we can also modify the weighting of the image condition during the inference stage:

$$\mathbf{Z}^{new} = \text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) + \lambda \cdot \text{Attention}(\mathbf{Q},\mathbf{K}',\mathbf{V}'). \tag{8}$$

where λ is weight factor, and the model becomes the original T2I diffusion model if $\lambda = 0$.

4. Experiment and Analysis

4.1. Painting-42

Collection. Despite the existence of numerous art datasets, such as VisualLink [57], Art500k [58], and Artemis [59], that facilitate AI learning, there is a scarcity of datasets specifically focused on Chinese paintings from distinct historical eras with unique styles or techniques. To advance the development of Chinese painting within the AI learning field, it is essential to construct more accurate and comprehensive datasets tailored to these specific contexts. To address this gap, we meticulously curated a dataset of 4,055 Chinese paintings spanning various historical periods and diverse artistic styles, sourced from online platforms and artist albums. The resolution distribution for these paintings is detailed in Figure 2. To ensure the stability and accuracy of the data, and to alleviate issues such as image blur, detail loss, distortion, or noise amplification, we standardized all paintings to a consistent resolution of 512×512. This resolution is consistent with the training dataset specifications of the SD model, ensuring model stability and output quality. Deviations from this resolution could potentially compromise the fidelity of the generated outputs. Throughout the standardization process, we placed significant emphasis on preserving the distinct painting styles, compositions, and aesthetic features characteristic of each historical era. Larger paintings were carefully cropped and segmented to ensure that each segment captured the essence of the original artwork. By doing so, we aimed to create a high-quality dataset that would facilitate the development of AI models capable of generating authentic and diverse Chinese paintings.

 doi:10.20944/preprints202408.1035.v1

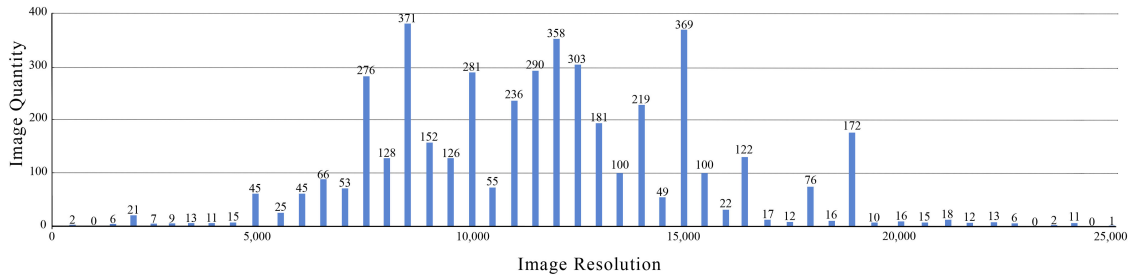

Figure 2. This table shows the size of the short edge of the images in the elemental data we collected, revealing that resolutions are predominantly concentrated between 400 and 3100 pixels. Notably, a significant number of these images exceed dimensions of 512 × 512 pixels.

Electronic replicas of Chinese paintings often fail to capture the intricate details and nuances found in the originals. To address this limitation, we meticulously adjusted the image parameters to enhance the painting features while minimizing noise. This process involved careful screening to remove low-resolution and redundant works, resulting in a curated collection of 4,055 high-quality Chinese painting images. These images cover various categories and styles of ancient and modern Chinese art, including blue-green landscapes, golden and blue-green landscapes, fine brush painting, flower and bird painting, etc. They also feature characteristic techniques such as plain drawing and texture strokes, showcasing the unique artistic traditions of Chinese painting. Each image was selected to ensure it authentically represents its respective era, enabling comparisons between painters and their unique styles or techniques. This curated dataset serves as valuable material for advancing machine learning research and facilitating the preservation and innovation of Chinese painting art in the digital age. Notably, the dataset includes works from 42 renowned Chinese artists that span seven different dynasties in Chinese history, providing a comprehensive representation of the evolution of Chinese painting styles. This initiative marks the creation of the first Chinese painting style dataset tailored specifically for T2I tasks. The distribution among dynasties within the dataset is visualized in Figure 3, illustrating the breadth and depth of the collection. By making this dataset available, we hope to contribute to the development of more sophisticated machine-learning models that can appreciate and generate Chinese paintings with greater accuracy and nuance.

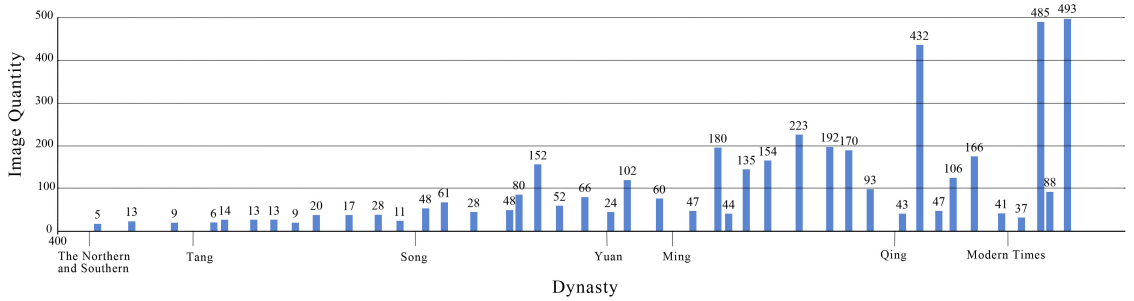

Figure 3. Painting-42 is classified based on the styles of 42 famous Chinese painting painters with historical influence, including artworks from six different dynasties, reflecting different style distributions.

Labels. To streamline the T2I generation process, we opted to extract elements and features of Chinese paintings from painting names and reannotate them using natural language, rather than annotating the paintings directly. The annotation workflow is illustrated in Figure 4. Initially, we used BLIP2 [60] for annotation and applied a filtering method (CapFilt) as an experimental approach. CapFilt leverages dataset guidance to filter out noisy and automatically generated image titles. However, after conducting a manual review and comparison, we identified areas where this method could be improved, particularly in accurately identifying free-hand or abstract expression techniques common in Chinese painting. To improve the quality of annotations, we consulted experts in Chinese art history

 doi:10.20944/preprints202408.1035.v1

and Chinese painting. We have improved annotations specifically for these types of images through detailed manual adjustments. Key enhancements included appending keywords such as "Chinese painting", "era", and "author" to each image description, highlighting core stylistic features of Chinese painting. Furthermore, we categorized distinct popular painting techniques associated with painters from different eras, ensuring the model could accurately distinguish and recognize unique expressive techniques characteristic of each era painter. Ultimately, we organized these annotated text-image pairs into a structured JSON file format for seamless integration into our T2I generation pipeline.

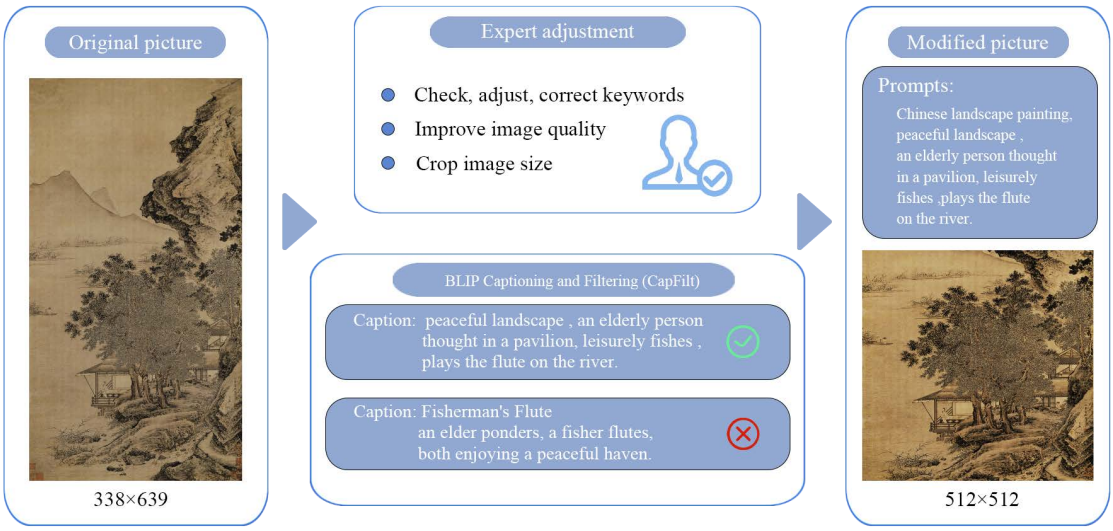

Figure 4. The image processing workflow for Painting-42 involves several steps. Initially, raw images underwent captioning using BLIP CapFit to generate adaptive subtitles, which were meticulously validated and corrected. Additionally, manual cropping and enhancement techniques were applied to the original images to achieve uniformly sized 512 × 512 images along with corresponding labels.

4.2. Quantitative Evaluation

Our objective in assisted design is to ensure that the Chinese paintings generated by our model exhibit harmonious layouts, vibrant colors, delicate brushwork, and a cohesive composition that balances diversity with stability, all guided by the provided theme description. To achieve this goal, we have chosen metrics that primarily assess the coherence between text descriptions and the aesthetic quality of the generated images. To evaluate the alignment between text and images, we employ a suite of metrics, including Style Loss [61], CLIP Score [62], and LPIPS [63]. Style loss measures the disparity between the generated image and a reference image, providing insights into how well the style has been captured. CLIP Score offers a comprehensive evaluation that considers factors such as color fidelity, texture, contrast, and clarity, translating these evaluations into numerical scores to assess overall image quality. LPIPS utilizes feature extraction and similarity calculation to provide a nuanced evaluation of image quality and similarity, utilizing deep learning techniques to accurately assess visual fidelity. By combining these metrics, we can rigorously evaluate our model's performance in aligning with textual prompts and producing aesthetically pleasing Chinese paintings that adhere to predefined thematic descriptions. This multi-faceted evaluation approach enables us to refine our model and ensure that the generated paintings meet the desired standards of coherence, aesthetic quality, and thematic consistency.

We conducted a comprehensive comparative analysis of our PDANet method against state-of-the-art models, including Midjourney and DALL-E 3, and presented empirical evidence to demonstrate our advances. To ensure a thorough evaluation across diverse categories and difficulty levels of text descriptions, we randomly selected eight prompts from PartiPrompts [64], a dataset comprising over 1,200 English prompts. For each prompt, we generated 50 painting-style images, and the final

 doi:10.20944/preprints202408.1035.v1

scores were averaged across all datasets to provide a robust assessment. The results, summarized in Table 1, showcase PDANet's impressive performance metrics. Notably, PDANet achieved a Style Loss score of 0.0088, which is approximately one-tenth of DALL-E 3's score, indicating a significant improvement in style capture. Furthermore, PDANet attained a CLIP Score of 0.8642, the highest among all evaluated methods, demonstrating its exceptional ability to generate images that align with the input text descriptions. In terms of LPIPS score, PDANet scored 0.6530, substantially surpassing other models and underscoring its strength in producing visually coherent images. These findings collectively underscore PDANet's remarkable capability to align generated images with input text descriptions, confirming its state-of-the-art status in producing painting-style images. The empirical evidence presented in this study validates the effectiveness of our proposed approach and highlights its potential for applications in T2I synthesis.

Table 1. Comparison of different generated images based on Style Loss, CLIP Score, and LPIPS.

Generated Image	Style Loss ↓	CLIP Score ↑	LPIPS ↓
DALL-E 3	0.0863	0.8074	0.7614
Reference Method	0.0514	0.8287	0.6770
Depth+Reference Method	0.0175	0.6326	0.7191
Midjourney	0.0121	0.8332	0.6994
PDANet(Ours)	**0.0088**	**0.8642**	**0.6530**

4.3. Qualitative Analysis

In this study, our primary objective was to extract key cue words from the PartiPrompts corpus to generate images with various Chinese painting styles, which we then applied to various models to evaluate their performance. To establish a benchmark for T2I conversion, we first developed a baseline model using SD. Furthermore, we augmented other models by incorporating a supplementary dataset of Chinese painting images, aiming to enrich their understanding of the characteristic styles and elements of Chinese painting. This approach sought to refine the models' ability to produce images that accurately capture the nuances and artistic subtleties unique to Chinese painting, thereby enhancing their overall performance in generating authentic and aesthetically pleasing images.

Through a series of systematic experiments and analyses, as illustrated in Figure 5, we evaluated various methods to generate painting-style images. Method A1 relies solely on style reference images, which, although exhibiting some characteristics of Chinese painting styles, suffers from instability in the generated image content. Method A2 generates images based solely on depth maps, but struggles to effectively capture and transform the distinctive features of Chinese painting styles. Method A3 combines style reference maps and depth maps as dual conditions, but still faces limitations and instability in style conversion. In contrast, our proposed model framework takes a multi-modal approach, leveraging a combination of modalities to extract and simulate the key style elements of Chinese painting, including stroke features, visual rhythm, and color style. This approach demonstrates enhanced learning capabilities and improved generation stability compared to the aforementioned methods. By incorporating multiple modalities, our model is able to capture the complex and nuanced characteristics of Chinese painting styles, resulting in more accurate and aesthetically pleasing generated images.

 doi:10.20944/preprints202408.1035.v1

Figure 5. Visual comparison results between PDANet and single-modal or bimodal methods are conducted under specific conditions. Prompts serve as controls randomly selected from a partial prompt corpus, incorporating reference images and depth images as painting-style conditions.Where A1 represents the Reference method, A2 represents the Depth method, and A3 represents the Depth+Reference method.

As shown in Figure 6, we conducted a comprehensive evaluation of the model's performance by utilizing a standard reference image and assessing the generated image styles produced by various models in response to different prompt words. The results demonstrate that PDANet exhibits exceptional proficiency in interpreting diverse prompt instructions, particularly excelling with complex and lengthy descriptions. It consistently generates Chinese painting-style images that faithfully match the given instructions, maintaining high fidelity in visual style while achieving notable strides in content diversity, innovation, and stability. Notably, PDANet's performance is characterized by its ability to deliver consistently satisfactory results, showcasing its robustness and reliability in generating high-quality images that meet the desired standards. The model's capacity to accurately interpret and respond to varied prompt instructions underscores its potential for applications in T2I synthesis, where the ability to generate diverse and contextually relevant images is paramount.

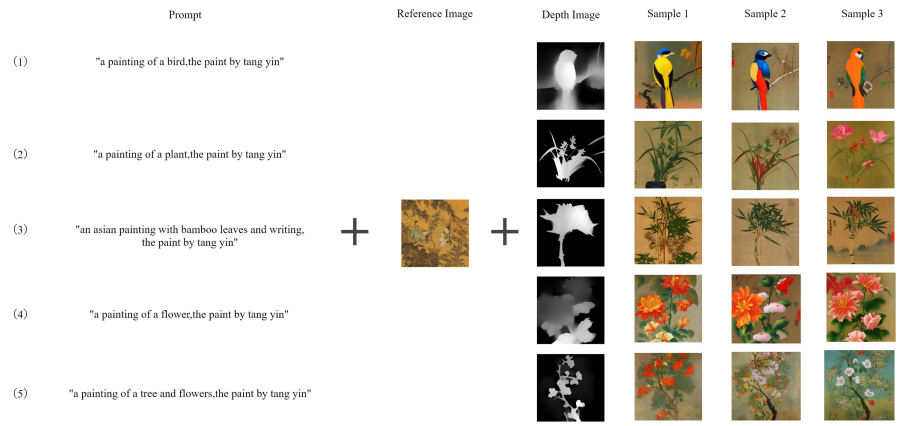

Figure 6. Using the same reference image as control variables and inputting different prompts, multimodal methods were compared with single-modal or bimodal methods primarily to assess the diversity and stability of the generated images.

This study conclusively demonstrates the superiority of PDANet in comprehending and interpreting deep-level cue words, thereby establishing a novel theoretical and practical foundation for

 doi:10.20944/preprints202408.1035.v1

stylization generation in the T2I field. By leveraging its advanced capabilities, PDANet sets a new benchmark for T2I synthesis, enabling the generation of highly stylized and contextually relevant images that accurately capture the essence of the input text.

4.4. User Study

In this user research evaluation, we conducted a comprehensive analysis of the images generated by the model across four key dimensions: visual-textual consistency, precision in capturing stylistic details, aesthetic preference, and stability of content and style. To assess the model's performance, users were asked to select the work that best aligned with these evaluation criteria from images generated by four different models. Visual-textual consistency was a crucial aspect of the evaluation, as it assesses the model's ability to maintain thematic coherence in design creation, similar to evaluating a designer's skill in this regard. This dimension forms a fundamental criterion for evaluating model performance, as it ensures that the generated images accurately reflect the content specified in the input text. By prioritizing visual-textual consistency, we can guarantee that the model produces images that are not only aesthetically pleasing but also contextually relevant and faithful to the original text.

Secondly, the assessment of learning style precision aims at evaluating the model's capability to faithfully replicate the style of a particular artist. This is achieved by measuring the similarity between the images generated by the model and the original works of the artist. In terms of aesthetic preference, a comprehensive quality assessment of the generated Chinese painting style images is conducted based on artistic principles, including composition, arrangement of elements, and visual expression. This dimension examines whether the model can produce visually appealing images while adhering to fundamental artistic principles.

Finally, in evaluating the stability of content and style, we focus on striking a balance between diversity and consistency in generating imaginative images with distinct styles. Our experimental results reveal that incorporating reference images as control conditions tends to constrain the model's creativity, often compromising the delicate balance between diversity and stability in the generated images. To comprehensively assess the performance of PDANet in terms of style diversity and stability, we conducted a comparative analysis with state-of-the-art models, including Midjourney, DALL-E 3, and SD. This evaluation enables us to rigorously examine the strengths and weaknesses of PDANet in achieving a harmonious balance between diversity and stability.

Our research sample includes 52 survey questionnaires from 12 cities in China, among which 35 participants have art education backgrounds and a profound appreciation for Chinese painting art. This ensured the evaluation's professionalism and reliability. As illustrated in Figure 7, we labeled the generated images from four models as follows: A (Midjourney), B (DALL-E 3), C (SD), and D (PDANet). In Case 1, the image produced by PDANet more accurately captured the essence of the keywords provided, while Models A, B, and C exhibited varying degrees of deviation in replicating the art style. This disparity highlights the superiority of PDANet in terms of style fidelity. In Case 2, we assessed each model's capacity to learn and reproduce the style inspired by the works of the renowned Ming Dynasty painter Tang Yin. Characterized by exquisite elegance and delicate, serene brushwork, Tang Yin's paintings are a hallmark of Chinese art. Notably, PDANet demonstrated a high level of consistency with Tang Yin's style in terms of visual aesthetics and brushstroke quality, underscoring its ability to capture the nuances of artistic expression.

 doi:10.20944/preprints202408.1035.v1

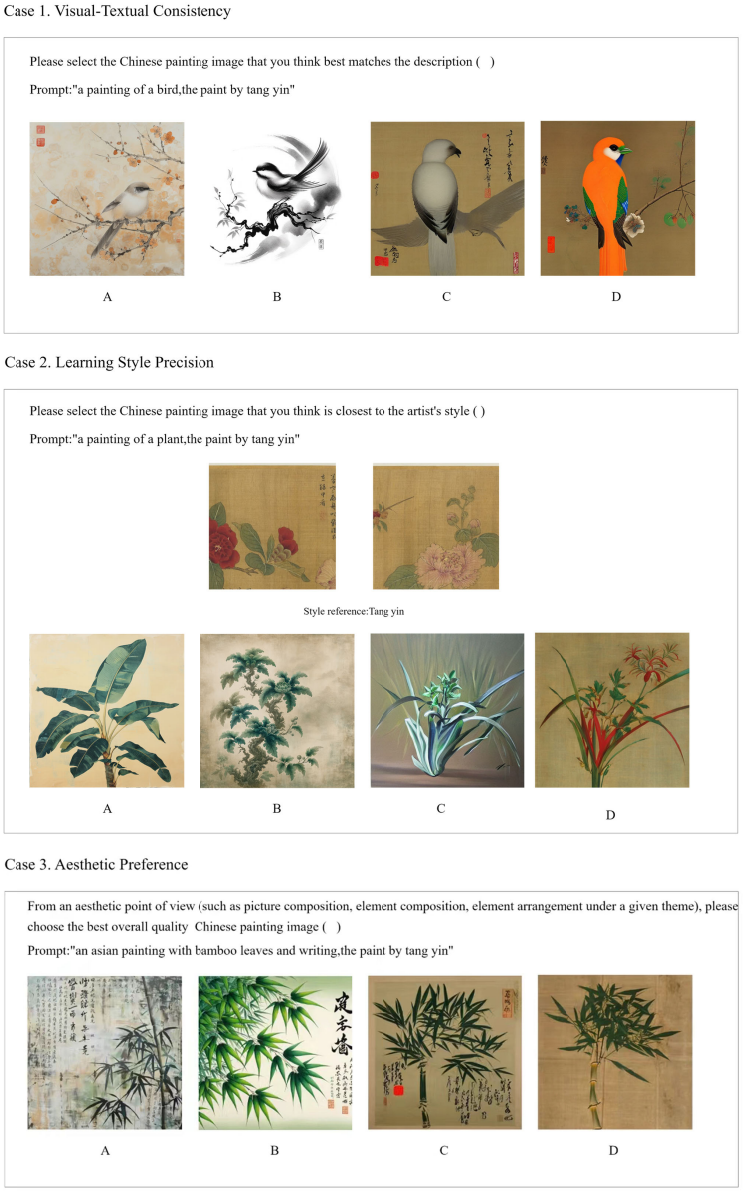

Figure 7. The Figure presents questionnaire cases evaluating visual-textual consistency, style precision learning, and aesthetic preference. In these cases, A represents Midjourney, B stands for DALL-E 3, C depicts SD, and D represents PDANet, our proposed method.

In the assessment of aesthetic preferences, Case 3 exemplifies that Option D consistently maintains a cohesive and visually appealing expression throughout the creative process, demonstrating a profound interpretation of the prompt "an Asian painting with bamboo leaves". This option aligns closely with the aesthetic expectations of the respondents and outperforms the alternatives in terms of artistic merit. As illustrated in Figure 8, the statistical data reveal that PDANet has garnered significant user preference and widespread recognition for its graphical consistency, stylistic precision, and aesthetic appeal. The results indicate that PDANet's output is considered more visually appealing and style-consistent than other options, emphasizing its ability to meet the aesthetic expectations of respondents.

 doi:10.20944/preprints202408.1035.v1

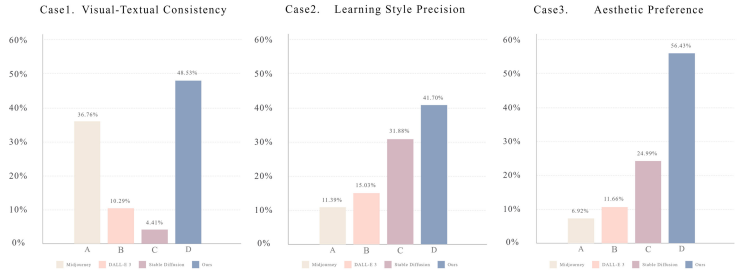

Figure 8. The user study comparing PDANet with state-of-the-art methods reveals user preferences. Here, we introduce three metrics for evaluating generative models in design tasks: visual-textual consistency, style precision learning, and aesthetic preference. The questionnaire content corresponding to these metrics is detailed in Figure 7.

In Case 4, illustrated in Figure 9, we conducted a comprehensive analysis of the style effects produced by each model when generating diverse images, with a particular emphasis on the stability and diversity of content and style. To evaluate style stability, we compared the performance of Midjourney (Option A), DALL-E 3 (Option B), SD (Option C), and PDANet (Option D) using the case depicted in the figure. The results reveal that option A exhibits instability in image composition and arrangement when generating Chinese painting styles, whereas option B fails to accurately capture the distinctive characteristics of Chinese painting from a specific period. Furthermore, Option C demonstrates limitations in style and expression, particularly in the stroke. In contrast, PDANet consistently delivers stable and high-quality results, effectively embodying the painting styles, compositions, and aesthetic nuances of images across different eras. As shown in Figure 10, the user preference results indicate that 53.93% of users preferred Option D, underscoring PDANet's exceptional performance in maintaining image style stability. This significant margin of preference demonstrates the superiority of PDANet in generating stable and visually appealing images that meet the aesthetic expectations of users.

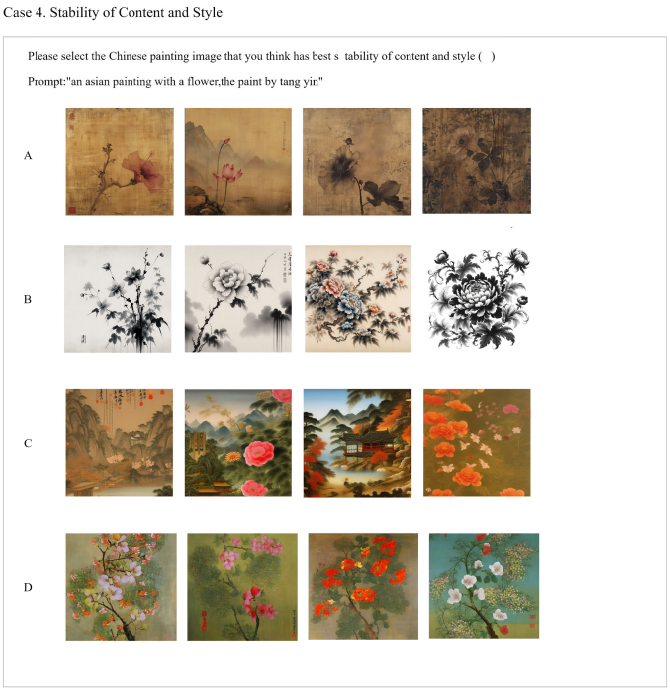

Figure 9. The above is a set of user interview questions used to evaluate the stability of the model. Here, A, B, C, and D represent Midjourney, DALL-E 3, SD, and PDANet, respectively. Our results demonstrate that our approach outperforms existing methods in terms of generating diverse and stable images, indicating its potential for improving T2I synthesis.

 doi:10.20944/preprints202408.1035.v1

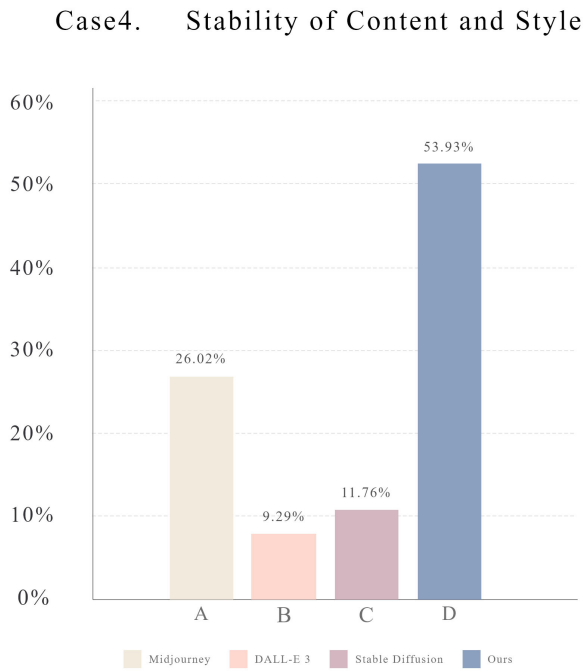

Figure 10. The questionnaire results on generative diversity demonstrated that PDANet exhibited a significant advantage in design richness, receiving 53.93% of the votes.

4.5. Generated Showcase

Figure 11 shows the Chinese paintings generated by the proposed method. These paintings have the painting-styles of Shen Zhou, Tang Yin, and Wang Hui. These generated works strictly follow the aesthetic principles of their respective styles, while achieving extremely high richness and refinement in detail expression. Specifically, the outline lines, the use of modeling techniques and the precise coloring of the colors in the picture are all displayed clearly and vividly, fully demonstrating the unique artistic style of each style and accurately capturing the core charm of traditional Chinese painting art.

Figure 11. Generated images in various styles by using the same depth image and reference images from different painters, following the method outlined in this paper. These images show a strong identity of the painter's style, and each style is unique and clear at a glance—for example, the painting styles of Shen Zhou, Tang Yin, and Wang Hui.

It is worth noting that although these three Chinese paintings were created based on the use of the same depth image, the final results generated are each unique, demonstrating excellent stylistic painting capabilities. The shape of the trees and the distribution of leaves in the picture are accurately depicted, forming a good visual hierarchy and sense of space. The successful application of this method marks its feasibility in practice. It not only reduces the technical difficulty in the creative

 doi:10.20944/preprints202408.1035.v1

process but also significantly improves the efficiency of creation, allowing designers to focus more on creativity and inspiration.

5. Conclusion

In conclusion, this paper underscores the crucial role of painting style and creativity in defining the essence and uniqueness of art, as these elements reflect the artist's personality and emotional tone, shaping audience perception, and enhancing the meaning and value of the work. The growing popularity of diffusion models in art design, animation, and gaming highlights their potential in original painting creation, poster design, and VI design. However, traditional creative processes are hindered by challenges such as slow innovation, high reliance on manual efforts, high costs, and limitations in large-scale replication. To address these challenges, we propose PDANet, a novel approach for style transformation that leverages the meticulously curated Painting-42 dataset, comprising 4,055 works by fifty-nine renowned Chinese painters from various periods. Using this data set, PDANet grasps the aesthetic intricacies of Chinese painting, providing users with rich design references. Furthermore, we introduce a lightweight Identity-Net for large-scale T2I models, which aligns internal knowledge with external control signals, enhancing the T2I model's capabilities. The trainable Identity-Net inputs image prompts into the U-Net encoder to generate new, diverse, and stable images. Our extensive quantitative experiments and qualitative analyses demonstrate that our approach surpasses current state-of-the-art methods, delivering high-quality generated content with broad applicability. This generative solution represents a significant advancement over traditional computer-aided design practices, offering a more efficient and innovative approach to creative design through deep learning. Using the power of AI, our approach has the potential to revolutionize the creative industry, enabling artists, designers, and developers to produce high-quality content with unprecedented ease and efficiency.

Author Contributions: Conceptualization, W.Y. and Y.Z.; methodology, Y.Z.; validation, Z.L., X.W. and Y.Q.; formal analysis, X.W. and Y.Q.; investigation, Z.L., X.W. and Y.Q.; data curation, Z.L., X.W. and Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.Z.; supervision, W.Y.; project administration, W.Y.; funding acquisition, W.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work is jointly supported by the Beijing Municipal High-Level Faculty Development Support Program (Grant No. BPHR202203072).

Data Availability Statement: The dataset used in this study will be made publicly available at a later stage and can be requested by contacting zhaoyifei@bigc.edu.cn.

Acknowledgments: The views and conclusions presented in this article are solely those of the authors and do not necessarily reflect the opinions of the sponsors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sun, Z.; Lei, Y.; Wu, X. Chinese Ancient Paintings Inpainting Based on Edge Guidance and Multi-Scale Residual Blocks. *Electronics* **2024**, *13*, 1212.
2. Li, H.; Fang, J.; Jia, Y.; Ji, L.; Chen, X.; Wang, N. Thangka Sketch Colorization Based on Multi-Level Adaptive-Instance-Normalized Color Fusion and Skip Connection Attention. *Electronics* **2023**, *12*, 1745.
3. Tu, Z.; Zhou, Q.; Zou, H.; Zhang, X. A multi-task dense network with self-supervised learning for retinal vessel segmentation. *Electronics* **2022**, *11*, 3538.
4. Anastasovitis, E.; Georgiou, G.; Matinopoulou, E.; Nikolopoulos, S.; Kompatsiaris, I.; Roumeliotis, M. Enhanced Inclusion through Advanced Immersion in Cultural Heritage: A Holistic Framework in Virtual Museology. *Electronics* **2024**, *13*, 1396.
5. Obradović, M.; Mišić, S.; Vasiljević, I.; Ivetić, D.; Obradović, R. The methodology of virtualizing sculptures and drawings: a case study of the virtual depot of the gallery of Matica Srpska. *Electronics* **2023**, *12*, 4157.
6. Gao, Y.; Wu, J. GAN-Based Unpaired Chinese Character Image Translation via Skeleton Transformation and Stroke Rendering. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, Vol. 34, pp. 722–729.

 doi:10.20944/preprints202408.1035.v1

7. Chen, N.; Li, H. Innovative Application of Chinese Landscape Culture Painting Elements in AI Art Generation. *Applied Mathematics and Nonlinear Sciences* **2024**.
8. Yi, M. Research on Artificial Intelligence Art Image Synthesis Algorithm Based on Generation Model. 2023 IEEE International Conference on Digital Image and Intelligent Computing (ICDIIME), 2023, pp. 116–121.
9. Zeng, W.; Zhu, H.l.; Lin, C.; Xiao, Z.y. A survey of generative adversarial networks and their application in text-to-image synthesis. *Electronic Research Archive* **2023**.
10. Yang, Q.; Bai, Y.; Liu, F.; Zhang, W. Integrated visual transformer and flash attention for lip-to-speech generation GAN. *Scientific Reports* **2024**, *14*.
11. He, Y.; Li, W.; Li, Z.; Tang, Y. GlueGAN: Gluing Two Images as a Panorama with Adversarial Learning. 2022 International Conference on Human-Machine Systems and Cybernetics (IHMSC), 2022.
12. Li, M.; Lin, L.; Luo, G.; Huang, H. Monet style oil painting generation based on cyclic generation confrontation network. *Journal of Electronic Imaging* **2024**, *33*.
13. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.
14. Borji, A. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586* **2022**.
15. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. International conference on machine learning. Pmlr, 2021, pp. 8821–8831.
16. Wu, Y.; Zhou, Y.; Xu, K. A Scale-Arbitrary Network for Chinese Painting Super-Resolution. 2023 ACM Symposium on Applied Computing, 2023.
17. Wang, W.; Huang, Y.; Miao, H. Research on Artistic Style Transfer of Chinese Painting Based on Generative Adversarial Network. 2023 International Conference on Artificial Intelligence and Information Technology (ACAIT), 2023.
18. Cheng, Y.; Huang, M.; Sun, W. VR-Based Line Drawing Methods in Chinese Painting. 2023 International Conference on Virtual Reality (ICVR), 2023.
19. Xu, H.; Chen, S.; Zhang, Y. Magical Brush: A Symbol-Based Modern Chinese Painting System for Novices. 2023 ACM Symposium on Applied Computing, 2023.
20. Yang, G.; Zhou, H. Teaching Chinese Painting Colour Based on Intelligent Image Processing Technology. *Applied Mathematics and Nonlinear Sciences* **2023**.
21. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
22. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125* **2022**.
23. Zhou, Y.; Wang, J.; Yang, Q.; Liu, H.; Zhu, J.Y. Image Synthesis with Latent Diffusion Models. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6794–6803.
24. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image Style Transfer Using Convolutional Neural Networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2414–2423.
25. Nikulin, Y.; Novak, R. Exploring the neural algorithm of artistic style. *arXiv preprint arXiv:1602.07188* **2016**.
26. Novak, R.; Nikulin, Y. Improving the neural algorithm of artistic style. *arXiv preprint arXiv:1605.04603* **2016**.
27. Li, C.; Wand, M. Precomputed real-time texture synthesis with markovian generative adversarial networks. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. Springer, 2016, pp. 702–716.
28. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.
29. Liu, Y.; Qin, Z.; Wan, T.; Luo, Z. Auto-painter: Cartoon image generation from sketch by using conditional Wasserstein generative adversarial networks. *Neurocomputing* **2018**, *311*, 78–87.
30. Zhao, H.; Li, H.; Cheng, L. Synthesizing filamentary structured images with GANs. *arXiv preprint arXiv:1706.02185* **2017**.

 doi:10.20944/preprints202408.1035.v1

31. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems* **2014**, *27*.
32. Xue, A. End-to-end chinese landscape painting creation using generative adversarial networks. Proceedings of the IEEE/CVF Winter conference on applications of computer vision, 2021, pp. 3863–3871.
33. Gatys, L.A.; Ecker, A.S.; Bethge, M. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* **2015**.
34. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **2021**, *34*, 8780–8794.
35. Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* **2021**.
36. Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; others. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems* **2021**, *34*, 19822–19835.
37. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; others. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* **2022**, *35*, 36479–36494.
38. Gafni, O.; Polyak, A.; Ashual, O.; Sheynin, S.; Parikh, D.; Taigman, Y. Make-a-scene: Scene-based text-to-image generation with human priors. European Conference on Computer Vision. Springer, 2022, pp. 89–106.
39. Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Zhang, Q.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; others. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324* **2022**.
40. Xue, Z.; Song, G.; Guo, Q.; Liu, B.; Zong, Z.; Liu, Y.; Luo, P. Raphael: Text-to-image generation via large mixture of diffusion paths. *Advances in Neural Information Processing Systems* **2024**, *36*.
41. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems* **2020**, *33*, 6840–6851.
42. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. International conference on machine learning. PMLR, 2015, pp. 2256–2265.
43. Song, Y.; Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* **2019**, *32*.
44. Kingma, D.; Salimans, T.; Poole, B.; Ho, J. Variational diffusion models. *Advances in neural information processing systems* **2021**, *34*, 21696–21707.
45. Esser, P.; Rombach, R.; Ommer, B. Taming transformers for high-resolution image synthesis. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12873–12883.
46. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; others. Learning transferable visual models from natural language supervision. International conference on machine learning. PMLR, 2021, pp. 8748–8763.
47. Feng, W.; He, X.; Fu, T.J.; Jampani, V.; Akula, A.; Narayana, P.; Basu, S.; Wang, X.E.; Wang, W.Y. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032* **2022**.
48. Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* **2022**.
49. Wang, C.; Tian, K.; Guan, Y.; Zhang, J.; Jiang, Z.; Shen, F.; Han, X.; Gu, Q.; Yang, W. Ensembling Diffusion Models via Adaptive Feature Aggregation. *arXiv preprint arXiv:2405.17082* **2024**.
50. Zhang, L.; Rao, A.; Agrawala, M. Adding conditional control to text-to-image diffusion models. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3836–3847.
51. Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; Wei, Y. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models. The Twelfth International Conference on Learning Representations, 2023.

 doi:10.20944/preprints202408.1035.v1

52. Wang, C.; Tian, K.; Zhang, J.; Guan, Y.; Luo, F.; Shen, F.; Jiang, Z.; Gu, Q.; Han, X.; Yang, W. V-Express: Conditional Dropout for Progressive Training of Portrait Video Generation. *arXiv preprint arXiv:2406.02511* **2024**.
53. Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; Tang, J. IMAGDressing-v1: Customizable Virtual Dressing. *arXiv preprint arXiv:2407.12705* **2024**.
54. Shen, F.; Ye, H.; Liu, S.; Zhang, J.; Wang, C.; Han, X.; Yang, W. Boosting Consistency in Story Visualization with Rich-Contextual Conditional Diffusion Models. *arXiv preprint arXiv:2407.02482* **2024**.
55. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer, 2015, pp. 234–241.
56. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1874–1883.
57. Seguin, B.; Striolo, C.; diLenardo, I.; Kaplan, F. Visual link retrieval in a database of paintings. Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I 14. Springer, 2016, pp. 753–767.
58. Mao, H.; Cheung, M.; She, J. Deepart: Learning joint representations of visual arts. Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1183–1191.
59. Achlioptas, P.; Ovsjanikov, M.; Haydarov, K.; Elhoseiny, M.; Guibas, L.J. Artemis: Affective language for visual art. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11569–11579.
60. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. International conference on machine learning. PMLR, 2022, pp. 12888–12900.
61. Ganguli, S.; Garzon, P.; Glaser, N. GeoGAN: A conditional GAN with reconstruction and style loss to generate standard layer of maps from satellite images. *arXiv preprint arXiv:1902.05611* **2019**.
62. Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R.L.; Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* **2021**.
63. Ghazanfari, S.; Garg, S.; Krishnamurthy, P.; Khorrami, F.; Araujo, A. R-LPIPS: An adversarially robust perceptual similarity metric. *arXiv preprint arXiv:2307.15157* **2023**.
64. Yu, J.; Xu, Y.; Koh, J.Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B.K.; others. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* **2022**, *2*, 5.