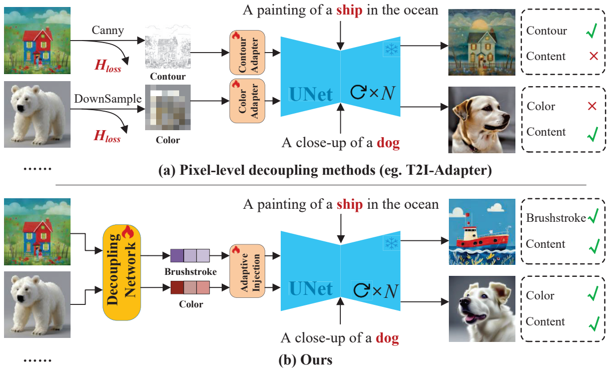
CO-PAINTER: Fine-Grained Controllable Image Stylization via Implicit Decoupling and Adaptive Injection

Bowen Fu[1] Wei Wei[1]* Jiaqi Tang[2] Jiangtao Nie[1] Yanyu Ye[1] Xiaogang Xu[3]
Ying-Cong Chen[2] Lei Zhang[1]
[1] Northwestern Polytechnical University
[2] The Hong Kong University of Science and Technology
[3] The Chinese University of Hong Kong
fubowen@mail.nwpu.edu.cn jtang092@connect.hkust-gz.edu.cn
xiaogangxu00@gmail.com yingcongchen@ust.hk
{weiweinwpu,nwpuzhanglei}@nwpu.edu.cn {niejiangtao,yyye_1}@mail.nwpu.edu.cn

Abstract

Controllable diffusion models have been widely applied in image stylization. However, existing methods often treat the style in the reference image as a single, indivisible entity, which makes it difficult to transfer specific stylistic attributes. To address this issue, we propose a fine-grained controllable image stylization framework, **CO-PAINTER**, *to decouple multiple attributes embedded in the reference image and adaptively inject them into the diffusion model. We first build a multi-condition image stylization framework based on the text-to-image generation model. Then, to drive it, we develop a fine-grained decoupling mechanism to implicitly separate the attributes from the image. Finally, we design a gated feature injection mechanism to adaptively regulate the importance of multiple attributes. To support the above procedure, we also build a dataset with fine-grained styles. It comprises nearly 48,000 image-text pairs samples. Extensive experiments demonstrate that the proposed model achieves an optimal balance between text alignment and style similarity to reference images, both in standard and fine-grained settings. Our code:* https://github.com/bowen310/Co-Painter



(a) Pixel-level decoupling methods (eg. T2I-Adapter)

(b) Ours

Figure 1. Comparison of different attribute decoupling strategies. (a) Pixel-level decoupling inevitably causes information loss, which leads to significant detail degradation. The contour branch generated inconsistent target image content (ship → house), while the color branch produced a color theme that did not align with the reference. (b) The implicit decoupling strategy using the entire image input achieves a superior transfer of attribute details.

1. Introduction

Image generation methods based on **C**ontrollable **D**iffusion **M**odels (CDMs) [16, 21, 22] have achieved significant breakthroughs across various domains, such as image stylization [3, 8, 36], image editing [11, 13, 26], and video generation [12, 14]. Among these applications, image stylization is dedicated to synthesizing highly artistic and expressive images that conform to the constraints of a given text or reference image [19]. Due to its potential in art creation, it has garnered increasing attention in recent years.

In practical applications, users always reference a variety of visual elements for their creations to meet specific visual effects or cultural expression demands. Consequently, when utilizing generative models, achieving fine-grained control over image details and style elements gradually becomes an emerging topic in the field of image stylization.

To address this issue, previous works [6, 10, 31, 36] have typically relied on text inversion or fine-tuning strategies to control style. However, the coarse text tokens learned in text inversion methods struggle to handle the transfer of specific style information from the reference images. Fine-tuning-based methods [28, 30] employ a lightweight network to transfer the detailed styles of additional reference images, achieving superior stylization performance. However, these methods are challenging to correctly decouple multiple vi-

*Corresponding author.

sual attributes because they treat the style in the reference image as a single, indivisible entity. To decouple image attributes, some methods [15] employ pixel-level strategies to separate image features (Figure 1-a). Although these models are effective at controlling structural attributes, like contours and layout, they still struggle to decouple and transfer fine-grained style attributes (such as brushstroke or color). These pixel-level operations often lead to significant information loss and conflict between textual and visual prompts.

To address these challenges, we propose **Co-Painter** for fine-grained controllable image stylization. Our goal is to transfer specific style attributes (brushstroke and color) from reference images to the target image while enabling users to precisely and flexibly control the generated style. To achieve this, we first construct a multi-condition image stylization framework based on a text-to-image generation model. This framework is designed to gradually generate the target image with the guidance of multiple conditions (Sec 4.1). To drive it, we then develop a fine-grained decoupling mechanism that implicitly decomposes attributes from the reference image. These decoupled embeddings can provide precise guidance in the diffusion process (Sec 4.2). Finally, we propose a gated feature injection mechanism to adaptively regulate the importance of multiple conditions across different layers (Sec 4.3). This facilitates effective collaboration between multiple conditions.

We also built a fine-grained style dataset that includes nearly 48,000 high-quality image-text pair samples. During training, we randomly select an image-text pair as the target image and prompt text, and then randomly assign three images with consistent content, brushstroke, and color as references. Extensive experiments show that compared to other methods, **Co-Painter** achieves an optimal balance between text alignment and style similarity to reference images in both standard and fine-grained settings.

Our contributions are as follows:

- We propose a fine-grained controllable image stylization model, **Co-Painter**. It can precisely decouple multiple visual attributes and adaptively inject them into the model, enhancing flexibility in artistic expression.
- We also build a dataset, comprising extensive high-quality image-text pairs in different brushstrokes and color schemes. This dataset can provide data support for fine-grained controllable image stylization.
- Extensive experiments demonstrate the effectiveness of the proposed method. It achieve superior qualitative and quantitative evaluation results than existing methods.

2. Related Work

Controllable Image Generation Recently, the controllability of the diffusion models [17] has gradually become a focal topic in the field of image generation [2, 9, 10]. Some methods [16, 21, 22] have implemented cross-attention modules to inject text embedding during the reverse process, enabling text-controlled image generation.

However, the text description is hard to provide precise guidance for those tasks requiring detailed structure control of image [33]. To address this issue, several studies [21, 23, 33] have introduced additional image conditions to enhance the structural controllability of generative models. ControlNet [33] adds various spatial control conditions to pre-trained text-to-image diffusion models, such as edges, depth, segmentation, and human poses, to further improve the control capabilities of diffusion models. Recently, T2I-Adapter [15] achieves the transfer of contour, color, and other information by utilizing edges, degraded images, and similar inputs. Although these methods provide various **structural control** information for image generation, they are still difficult to decouple and fine-grained transfer the style attributes of the reference image.

Therefore, our research centers on developing a controllable image customization framework to address the challenges mentioned above. Our goal is to provide precise visual guidance for **fine-grained style attribute transfer** through the use of multiple image prompts.

Image Stylization Image stylization aims to generate high-fidelity artistic images that conform to the requirements of text or image prompts. Traditional methods have employed techniques such as deep convolutional networks [7], Transformers [5], and GANs [29], driving substantial progress in image stylization. In recent years, considering the remarkable potential of controllable diffusion models, various studies have been conducted to explore their performance in high-quality, realistic stylized image synthesis tasks. These methods primarily use text inversion [4, 35, 36] or fine-tuning [32] techniques to control the style.

However, since text embeddings cannot provide detailed guidance for image generation, text inversion-based methods often face significant challenges in transferring special style attributes in images [30]. In contrast, fine-tuning-based methods excel at capturing detailed stylistic information from image prompts. Recently, DEADiff [19] and ArtAdapter [3] utilized various feature injection mechanisms to mitigate the issue of content leakage in some early fine-tuning-based studies [15, 32], achieving better image stylization. However, the fixed injection strategy makes it difficult to adapt the variations across different cross-attention layers. Furthermore, these methods often focus on the **entire style transfer** of the reference image, neglecting the fine-grained control of individual style attributes.

To address these challenges, our paper introduces a fine-grained decoupling and a gated feature injection mechanism. These modules offer precise style guidance and adaptive injection into the diffusion model's layers for **fine-grained style attribute transfer**.

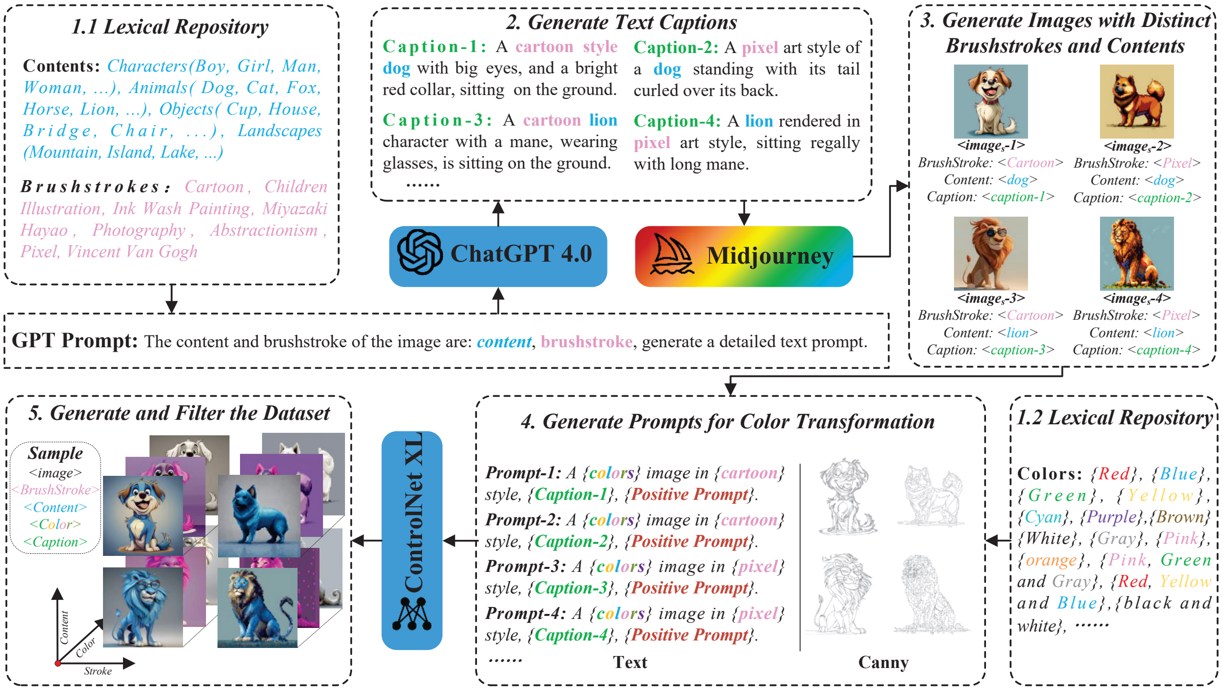

1.1 Lexical Repository
Contents: Characters(Boy, Girl, Man, Woman, ...), Animals(Dog, Cat, Fox, Horse, Lion, ...), Objects(Cup, House, Bridge, Chair, ...), Landscapes (Mountain, Island, Lake, ...)
Brushstrokes: Cartoon, Children Illustration, Ink Wash Painting, Miyazaki Hayao, Photography, Abstractionism, Pixel, Vincent Van Gogh
2. Generate Text Captions
Caption-1: A cartoon style dog with big eyes, and a bright red collar, sitting on the ground.
Caption-2: A pixel art style of a dog standing with its tail curled over its back.
Caption-3: A cartoon lion character with a mane, wearing glasses, is sitting on the ground.
Caption-4: A lion rendered in pixel art style, sitting regally with long mane.
......
ChatGPT 4.0
Midjourney
3. Generate Images with Distinct Brushstrokes and Contents
<image$_s$-1> BrushStroke: <Cartoon> Content: <dog> Caption: <caption-1>
<image$_s$-2> BrushStroke: <Pixel> Content: <dog> Caption: <caption-2>
<image$_s$-3> BrushStroke: <Cartoon> Content: <lion> Caption: <caption-3>
<image$_s$-4> BrushStroke: <Pixel> Content: <lion> Caption: <caption-4>
GPT Prompt: The content and brushstroke of the image are: content, brushstroke, generate a detailed text prompt.
5. Generate and Filter the Dataset
Sample
<image>
<BrushStroke>
<Content>
<Color>
<Caption>
Content
Color
Stroke
ControlNet XL
4. Generate Prompts for Color Transformation
Prompt-1: A {colors} image in {cartoon} style, {Caption-1}, {Positive Prompt}.
Prompt-2: A {colors} image in {cartoon} style, {Caption-2}, {Positive Prompt}.
Prompt-3: A {colors} image in {pixel} style, {Caption-3}, {Positive Prompt}.
Prompt-4: A {colors} image in {pixel} style, {Caption-4}, {Positive Prompt}.
......
Text
Canny
1.2 Lexical Repository
Colors: {Red}, {Blue}, {Green}, {Yellow}, {Cyan}, {Purple},{Brown} {White}, {Gray}, {Pink}, {orange}, {Pink, Green and Gray}, {Red, Yellow and Blue},{black and white}, ...

Figure 2. The details of the dataset construction. **1)** We created a lexical repository with various terms. **2)** These terms are randomly combined to generate image captions. **3)** Then, images with varying brushstrokes and contents are acquired. **4)** Additionally, we combine color terms and canny maps for color transformations. **5)** Finally, after filtering and checking, we created a fine-grained stylized dataset.

3. Building Dataset

Existing text-image datasets, such as LAION-5B [25] and LAION-400M [24], contain a large amount of high-quality image-text pairs. However, due to the lack of large-scale paired samples with similar image attributes, they are limited in addressing the challenges posed by this task. Although WikiArt [18] partitions image data based on similar brushstrokes, it still lacks paired samples with consistent color schemes. Hence, to achieve fine-grained control over style attributes, there is an urgent need to construct a new dataset that fulfills our requirements.

Overview To address the scarcity of paired samples with consistent image attributes (content, brushstroke, and color), we constructed a novel dataset utilizing GPT-4 [1], Midjourney[1], and ControlNet [33]. The data format can be represented as follows,

$$< (\mathrm{T_{tgt}} \mid \mathrm{I_{tgt}}) \Rightarrow (\mathrm{I_{con}}, \mathrm{I_{bst}}, \mathrm{I_{col}}) >, \tag{1}$$

where $\mathrm{I_{tgt}}$ and $\mathrm{T_{tgt}}$ represent the target image and the corresponding prompt text, respectively. $\mathrm{I_{con}}$, $\mathrm{I_{bst}}$, and $\mathrm{I_{col}}$ denote the reference images that are consistent with the target image in terms of content, brushstroke, and color, respectively. Figure 2 shows the detailed pipeline of the dataset building.

Construct Lexical Repository To meet the above requirement, a lexical repository was built. First, we gather 8 distinct terms to describe the brushstrokes, including "Cartoon," "Children Illustration," "Ink wash Painting," "Oil Painting," "Pixel," "Miyazaki Hayao," "Photo," and "Vincent Van Gogh." Second, we collect more than 20 color schemes to describe color attributes, such as "red", "green", "black, and white". Finally, we collect approximately 130 common terms to describe the content in images. These terms cover a wide range of subjects, including characters, animals, objects, and landscapes. These terms will be randomly combined to generate detailed image captions.

Create Image Captions We used GPT-4 [1] to create detailed text captions for images. To begin with, we prepared 10 prompt templates to provide instructions to GPT-4 [1]. Subsequently, one of these templates was randomly selected to combine with content and brushstroke terms randomly drawn from the lexical repository. As a result, nearly 3,000 detailed image captions were generated ($\mathrm{T_{tgt}}$). This strategy was able to enhance data diversity and avoid the generation of similar or repetitive image samples.

Generate Images with Various Contents & Brushstrokes After obtaining a detailed description of the image, we used Midjourney v6.0 to perform high-quality image synthesis. First, the text captions were input into the Midjourney to synthesize 4 image samples. Next, the samples that aligned with input instructions were up-sampled to the resolution of 1024 × 1024. Finally, following manual checking and categorization, a high-quality image set with diverse brushstrokes was constructed. In total, we generated 2290 high-quality images ($\mathrm{I_{tgt}}$). A large number of paired images with consistent brushstrokes ($\mathrm{I_{bst}}$) or content ($\mathrm{I_{con}}$) can be found

[1] https://www.midjourney.com

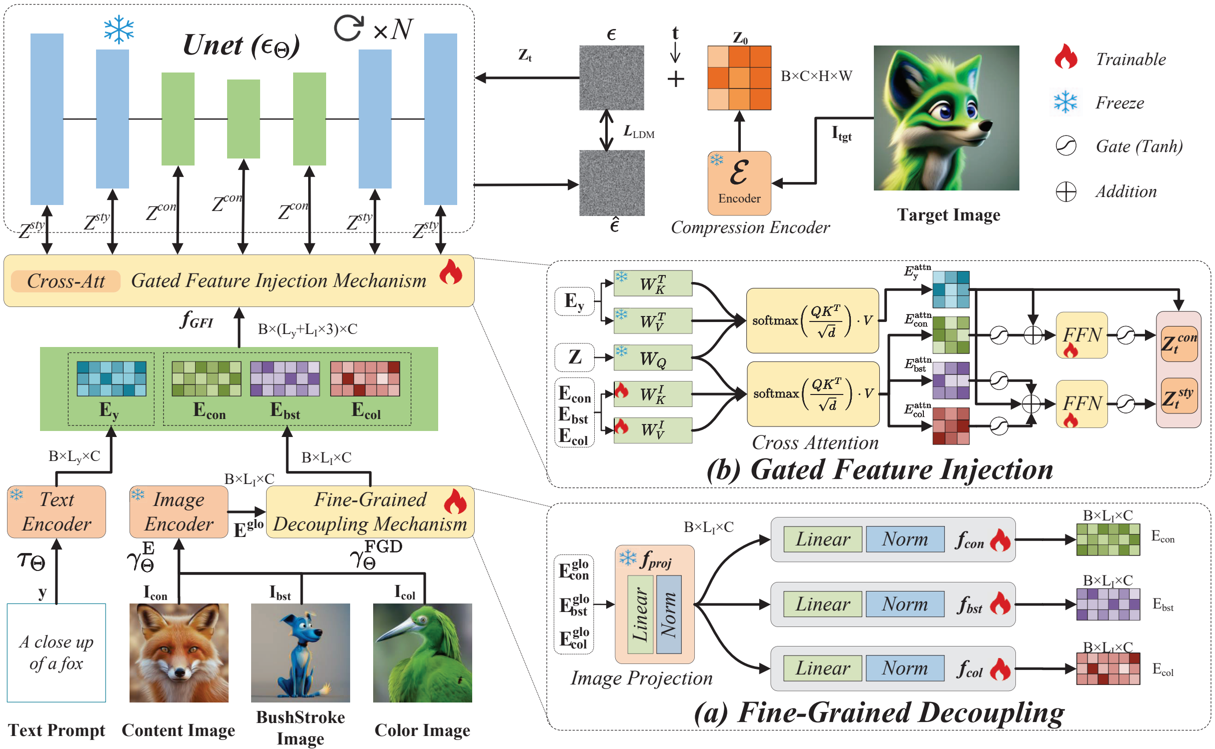


Figure 3. The overall structure of CO-PAINTER. We introduce a fine-grained decoupling strategy and a gated feature injection mechanism into the text-to-image model to achieve fine-grained controllable image stylization.

within these samples.

Create Prompts for Color Transformation Although the previous data acquisition strategy generates numerous paired images with consistent brushstroke and content attributes, it still lacks paired images (I_{col}) with consistent color attributes. To address this issue, different color terms were randomly combined with the original samples' brushstroke terms and captions to create text prompts. Besides, we compute the canny map for each sample to provide the text-to-image model with strong structural priors, ensuring that the overall structure of the image remains unchanged.

Generate Images with Distinct Colors Driven by the above prompts, we introduce ControlNet [33] to generate high-quality images with varying color attributes. The structure of our dataset can be represented in a 3-dimensional coordinate system (Figure 2-5). Once a target image is randomly selected, numerous samples with matching content, brushstroke, and color attributes can be obtained within the dataset. In the data construction process, we combined CLIP [20] with manual supervision to filter and check the synthesized images, ensuring data accuracy and diversity. Ultimately led to the filtering of approximately 20% of anomalous data.

4. Methodology

Summary Our method aims to stylize images by utilizing different image attributes at fine granularity. To tackle this challenge, we propose an end-to-end framework, **CO-PAINTER**, to decouple different image attributes implicitly and then adaptively inject them into the diffusion model (Figure 3). In Sec. 4.1, we first extend the text-to-image model and construct a multi-conditional image stylization framework. To drive it, we establish a **F**ine-**G**rained **D**ecoupling (FGD) mechanism in Sec. 4.2 to separate multiple image attributes. Finally, we develop a **G**ated **F**eature **I**njection (GFI) mechanism in Sec. 4.3 to control the importance of multiple conditions for injecting attributes.

4.1. Diffusion Model with Multiple Conditions

Fine-grained controllable image stylization aims to effectively decouple the different attributes from the reference image and utilize these decoupled representations as guiding conditions for generating stylized images.

Referring to previous work, such as ControlNet [33], we can construct a diffusion model with multiple conditions for the backbone of our problem. So, assuming the input of the diffusion model is $(\mathrm{y}, \mathrm{I}_1, \ldots, \mathrm{I}_N)$, where y is the text description, and $\mathrm{I}_i, i \in [1, N]$ represents reference images. We employ a forward process $q\left(\mathbf{z}_{1:T} \mid \mathbf{z}_0\right)$ to progressively add noise ϵ to the image features z_0 in the latent space,

$$q\left(\mathbf{z}_{1:T} \mid \mathbf{z}_0\right) = \prod_{t=1}^{T} \mathcal{N}\left(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t \mathbf{I}\right). \quad (2)$$

Our goal is to utilize the reverse process $p_\Theta(z_{0:T})$ to progressively generate the target image I_{tgt} with multiple conditions $c = (c_y, c_1, ..., c_N)$, and $c_y = \tau_\Theta(y)$, $c_i = \gamma_\Theta(I_i)$,

$$p_\theta\left(\mathbf{z}_{0:T}\right) = p\left(\mathbf{z}_T\right) \prod_{t=1}^{T} \mathcal{N}\left(\mathbf{z}_{t-1}; \boldsymbol{\epsilon}_\theta\left(\mathbf{z}_t, c, t\right), \boldsymbol{\Sigma}_\theta\left(\mathbf{z}_t, t\right)\right), \quad (3)$$

where $p\left(\mathbf{z}_T\right) = \mathcal{N}\left(\mathbf{z}_T; \mathbf{0}, \mathbf{I}\right)$, ϵ_θ refers to the denoising model. $\tau_\Theta(\cdot)$ is the text encoder, $\gamma_\Theta(\cdot)$ is the encoder for input images.

Finally, the denoising model ϵ_Θ predicts the noise in the latent features at the time step t based on this control information c, optimizing the process using MSE loss, as

$$L_{\mathrm{LDM}} = \mathbb{E}_{\mathcal{E}(\boldsymbol{x}_0), \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \boldsymbol{c}, t} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\boldsymbol{z}_t, \boldsymbol{c}, t\right) \right\|_2^2. \quad (4)$$

4.2. Decoupling Image Attributes

Based on the initial structure above, we can integrate multiple conditions. However, it is challenging to accurately use one single attribute as control information to generate stylized images due to the high coupling of these attributes. To solve this problem, most previous studies only considered coarse-grained **content** and **style** attributes in images while ignoring fine control over more fine-grained style attributes, making it difficult to match the desired attributes to achieve accurate image generation.

To address the above problem, we first finely decouple the **content**, **brushstroke**, and **color** attributes from reference images (Figure 3-a). Therefore, in our image encoder $\gamma_\Theta(\cdot)$, we include the original image encoder $\gamma_\Theta^{\mathrm{E}}$, and then propose the fine-grained decoupling mechanism $\gamma_\Theta^{\mathrm{FGD}}$, for implicitly decoupling distinct style attribute embeddings ($\mathrm{E}_{\mathrm{con}}$, $\mathrm{E}_{\mathrm{bst}}$, $\mathrm{E}_{\mathrm{col}}$) from multiple reference images ($\mathrm{I}_{\mathrm{con}}$, $\mathrm{I}_{\mathrm{bst}}$, $\mathrm{I}_{\mathrm{col}}$), as,

$$\begin{cases} \mathrm{E}_{\mathrm{con}} & = \gamma_\Theta^{FGD}(\gamma_\Theta^{E}(\mathrm{I}_{\mathrm{con}})) \\ \mathrm{E}_{\mathrm{bst}} & = \gamma_\Theta^{FGD}(\gamma_\Theta^{E}(\mathrm{I}_{\mathrm{bst}})) \\ \mathrm{E}_{\mathrm{col}} & = \gamma_\Theta^{FGD}(\gamma_\Theta^{E}(\mathrm{I}_{\mathrm{col}})), \end{cases} \quad (5)$$

where $\gamma_\Theta^{E}(\cdot)$ is image encoder, and γ_Θ^{FGD} is define as,

$$\gamma_\Theta^{FGD}(\mathrm{E}_\delta^{\mathrm{glo}}) = f_\delta(f_{\mathrm{proj}}(\mathrm{E}_\delta^{\mathrm{glo}})), \ \delta \in \{\mathrm{con}, \mathrm{bst}, \mathrm{col}\}, \quad (6)$$

here $\mathrm{E}_\delta^{\mathrm{glo}}$ indicates the global feature embedding of the input content image, brushstroke image, or color image. $f_{\mathrm{proj}}(\cdot)$ is the shared image projection. $f_\delta(\cdot)$ is used to separate different attributes.

Training Procedure During training, we select 3 reference images per iteration that align with the target image in content, brushstroke, and color to provide implicit feature guidance. This strategy enables the three decoupling structures to distill the implicit representations of distinct image attributes from the reference images.

4.3. Adaptively Injecting Attributes

Despite now we can effectively separate visual attributes, it is still hard to combine multiple conditions and inject them into the diffusion model. Previous approaches [19, 32] usually employ techniques like addition or feature concatenation to integrate textual and visual conditions. Yet, this fixed injection method often fails to appropriately balance the influence of various attributes within the diffusion model.

To address the problem above, we propose a gated feature injection mechanism (Figure 3-b), as

$$\begin{aligned} Z^{\mathrm{sty}}, Z^{\mathrm{con}} &= f_{\mathrm{GFI}}(\mathrm{E}_{\mathrm{y}}, \mathrm{E}_{\mathrm{con}}, \mathrm{E}_{\mathrm{bst}}, \mathrm{E}_{\mathrm{col}}) \\ &= f_{\mathrm{g}}[f_{\mathrm{attn}}^{y}(\mathrm{E}_{\mathrm{y}}), f_{\mathrm{attn}}^{i}(\mathrm{E}_{\mathrm{con}}, \mathrm{E}_{\mathrm{bst}}, \mathrm{E}_{\mathrm{col}})], \end{aligned} \quad (7)$$

where $f_{\mathrm{g}}(\cdot)$ represents the feature fusion function with a learnable gate and feed-forward network, $Z^{\mathrm{sty}}, Z^{\mathrm{con}}$ refer to the latent features that are inserted into the fine and coarse layers of the diffusion model, respectively. $f_{\mathrm{attn}}^{y}(\cdot)$ and $f_{\mathrm{attn}}^{i}(\cdot)$ denote the cross-attention for text or image features respectively, which is defined as,

$$f_{\mathrm{attn}}(Q, K, V) = \mathtt{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V. \quad (8)$$

For the text attention, we use the text embeddings E_{y} to get K and V, and the internal features z of the denoising model to get Q, to compute the text-related embeddings $\mathrm{E}_{\mathrm{y}}^{\mathrm{attn}}$ as,

$$\begin{cases} \mathrm{E}_{\mathrm{y}}^{\mathrm{attn}} = f_{\mathrm{attn}}^{\mathrm{y}}(Q_{\mathrm{T}}, K_{\mathrm{T}}, V_{\mathrm{T}}) \\ Q_{\mathrm{T}} = W_{\mathrm{Q}} \cdot z, K_{\mathrm{T}} = W_{\mathrm{K}}^{\mathrm{T}} \cdot \mathrm{E}_{\mathrm{y}}, V_{\mathrm{T}} = W_{\mathrm{V}}^{\mathrm{T}} \cdot \mathrm{E}_{\mathrm{y}}, \end{cases} \quad (9)$$

where W_{Q}, $W_{\mathrm{K}}^{\mathrm{T}}$, and $W_{\mathrm{V}}^{\mathrm{T}}$ represent the frozen linear weights for the diffusion latent feature and text feature, respectively.

For the image attention, we use the three image embeddings $\mathrm{E}_{\mathrm{con}}$, $\mathrm{E}_{\mathrm{bst}}$, and $\mathrm{E}_{\mathrm{col}}$ as K and V, respectively, to compute the corresponding embeddings $\mathrm{E}_{\mathrm{con}}^{\mathrm{attn}}$, $\mathrm{E}_{\mathrm{bst}}^{\mathrm{attn}}$ and $\mathrm{E}_{\mathrm{col}}^{\mathrm{attn}}$ with the intermediate features z of the denoising model:

$$\begin{cases} \mathrm{E}_{\alpha}^{\mathrm{attn}} = f_{\mathrm{attn}}^{i}(Q_{\mathrm{I}}, K_{\mathrm{I}}, V_{\mathrm{I}}) \\ Q_{\mathrm{I}} = W_{\mathrm{Q}} \cdot z, K_{\mathrm{I}} = W_{\mathrm{K}}^{\mathrm{I}} \cdot \mathrm{E}_{\alpha}, V_{\mathrm{I}} = W_{\mathrm{V}}^{\mathrm{I}} \cdot \mathrm{E}_{\alpha} \\ \alpha \in \{\mathrm{con}, \mathrm{bst}, \mathrm{col}\}, \end{cases} \quad (10)$$

where $W_{\mathrm{K}}^{\mathrm{I}}$ and $W_{\mathrm{V}}^{\mathrm{I}}$ represent the shared linear layers for all image attributes. E_{α} denotes the image embeddings.

Finally, to fuse the different visual attributes into Diffusion, we propose a learnable gate for this process, as,

$$\begin{cases} Z^{\mathrm{sty}} = \mathrm{E}_{\mathrm{y}}^{\mathrm{attn}} \oplus g \cdot f_{\mathrm{FFN}}[g(\mathrm{E}_{\mathrm{bst}}^{\mathrm{attn}}) \oplus g(\mathrm{E}_{\mathrm{col}}^{\mathrm{attn}}) \oplus \mathrm{E}_{\mathrm{y}}^{\mathrm{attn}})] \\ Z^{\mathrm{con}} = \mathrm{E}_{\mathrm{y}}^{\mathrm{attn}} \oplus g \cdot f_{\mathrm{FFN}}[g(\mathrm{E}_{\mathrm{con}}^{\mathrm{attn}}) \oplus \mathrm{E}_{\mathrm{y}}^{\mathrm{attn}})], \end{cases} \quad (11)$$

where gate function $g(\cdot) = \mathrm{tanh}[(\cdot); \mathrm{T}]$, and T is a learnable temperature to control the magnitude of the activation function. $f_{FFN}(\cdot)$ is the feed-forward network.

Inspired by DEADiff [19], we selectively fuse the brushstroke embedding $\mathrm{E}_{\mathrm{bst}}^{\mathrm{attn}}$ and the color embedding $\mathrm{E}_{\mathrm{col}}^{\mathrm{attn}}$ into the layers with more local information to inject the style information, Z^{sty}. Besides, the content embedding $\mathrm{E}_{\mathrm{con}}^{\mathrm{attn}}$ is integrated into the layers with more global information to inject the content information, Z^{con}. Based on this, various conditions can be adaptively fused and injected into the

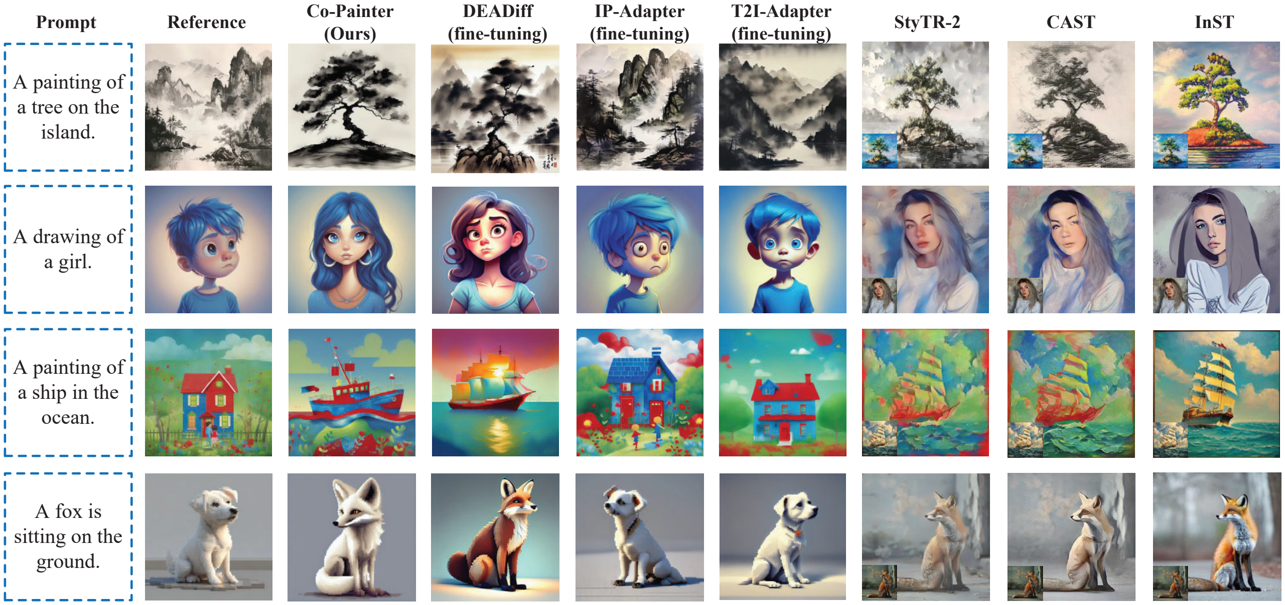

Prompt
Reference
Co-Painter (Ours)
DEADiff (fine-tuning)
IP-Adapter (fine-tuning)
T2I-Adapter (fine-tuning)
StyTR-2
CAST
InST
A painting of a tree on the island.
A drawing of a girl.
A painting of a ship in the ocean.
A fox is sitting on the ground.

Figure 4. Qualitative evaluation in a standard setting. The results demonstrate that despite **CO-PAINTER** being a fine-grained image stylization model, it exhibits superior style transfer and text alignment capabilities compared to the sota methods.

UNet. The learnable gates can flexibly adjust to different injection requirements, preventing information leakage and effectively transferring detailed attribute information.

5. Experiments

Training and Testing We utilized our dataset in Sec.3 for model training and testing. We randomly selected 90% of the samples as the train set and 10% as the test set. For the training, in each iteration step, we randomly selected a sample as the target image and then chose three reference images that matched it in content, brushstroke, and color for joint training. For the testing, we randomly selected 3 samples from the test set, serving as the content image, brushstroke image, and color image, respectively. The text prompt is the caption of the content image. As a result, 400 test tuples were constructed for evaluation.

Evaluation Metrics Inspired by DEADiff [19], we utilized the cosine similarity of CLIP [20] to evaluate **T**ext **A**lignment (TA) and **S**tyle **S**imilarity (SS). Moreover, to effectively evaluate the fine-grained controllable generation capabilities of **CO-PAINTER**, we constructed **CON**tent **A**lignment (CONA), **B**rush**ST**roke **S**imilarity (BSTS), and **COL**or **S**imilarity (COLS) using the text descriptions of content, brushstroke, and color from the three reference images. Additionally, we utilized the LAION-Aesthetics Predictor[2] to assess the average **I**mage **Q**uality (IQ) generated by each method.

Implementation Details Stable Diffusion 1.5 [22] is employed as the base model for this paper. We adopted the same coarse and fine layer division strategy as DEADiff [19]. The CLIP [20] (ViT-L/14) model was used as both the text and image encoders. The pre-trained IP-Adapter [32] was utilized to initialize the parameters of the image projection (f_{proj}) and the attention linear (W_K^I, W_V^I). During training, we used the AdamW optimizer with a batch size of 4, a learning rate of 1e-4, and 100,000 iterations. All experiments were conducted on 4 NVIDIA RTX 3090 GPUs. During testing, we employed the DDIM [27] with 30 steps for sampling.

Method	Type	IQ↑	SS↑	TA↑
InST [35]	Inversion	5.66	24.9	23.2
CAST [34]	Conventional	5.46	24.3	**25.9**
StyTR-2[5]	Conventional	5.48	24.8	24.5
T2I-Adapter [15]	Diffusion	6.02	**37.1**	13.5
IP-Adapter [32]	Diffusion	6.07	36.9	14.3
DEADiff [19]	Diffusion	6.09	26.4	24.9
CO-PAINTER(Ours)	Diffusion	**6.14**	27.8	24.5

Table 1. Quantitative evaluation in standard setting. Methods that excessively reference the style image tend to have a high SS, while those with higher TA struggle to capture detailed style information. **Balancing** SS and TA was a key goal for image stylization methods [3, 19].

5.1. Evaluation of Standard Image Stylization

Baselines We conduct a comprehensive comparison of the proposed **CO-PAINTER** with state-of-the-art (SOTA) methods in the standard image stylization setting. These methods include **1)** conventional methods (CAST [34] and StyleTr2 [5]), **2)** inversion based methods (InST [35]), as well as **3)** diffusion-based methods (T2I-Adapter [15], IP-Adapter [32], and DEADiff [19]). For fairness in comparison, we utilized Midjourney to generate content images for

[2] https://github.com/christophschuhmann/improved-aesthetic-predictor

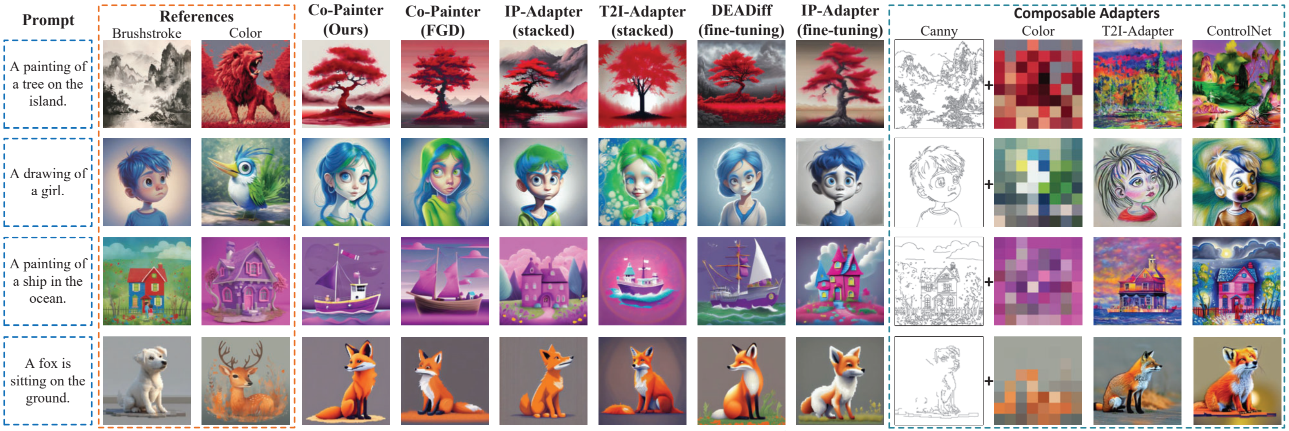


Figure 5. Qualitative evaluation in a fine-grained setting. The results show that **Co-Painter** excels in decoupling fine-grained style attributes and has superior adaptive feature injection capabilities.

CAST [34], StyleTr2 [5], and InST [35] using the same text prompts as the other methods.

Quantitative Evaluation Analysis Table 1 demonstrates the effectiveness of our model under the standard image stylization setting. Other baselines often struggle to capture fine-grained style and semantic information or overfit to the image prompt, resulting in suboptimal stylization performance. However, despite being specifically designed for fine-grained image stylization, **Co-Painter** achieves the optimal balance between SS and TA. Furthermore, through the collaborative interaction of the FGD and GFI modules, it attains state-of-the-art IQ.

Qualitative Evaluation Analysis Figure 4 demonstrates that the proposed method outperforms other baselines, achieving the best visual outcomes. Other methods exhibit varying degrees of visual degradation, such as inconsistent styles (DEADiff [19]), content leakage (IP-Adapter [32] and T2I-Adapter [15]), and loss of fine-grained details (StyTR-2 [5], CAST [34], and InST [35]). In contrast, **Co-Painter** precisely transfers detailed style information while maintaining controllability under textual conditions.

5.2. Evaluation of Fine-Grained Image Stylization

Baselines To assess the model's ability to transfer fine-grained style attributes, we expanded the reference image set from one to two, focusing on brushstroke and color. However, we found that there are no methods specifically designed for decoupling abstract brushstrokes and color after investigation. Therefore, we employ the method based on Canny edge and downsample color maps to illustrate the limitations of existing methods (T2I-Adapter [15] and ControlNet [33]). To enhance the comparability of our experiments, we also developed various comparative methods for implicitly decoupling brushstroke attributes from reference images based on existing techniques (DEADiff [19], IP-Adapter (fine-tuning & stacked) [32], T2I-Adapter (stacked) [15] and FGD in **Co-Painter**).

Quantitative Evaluation Analysis Table 2 demonstrates the superior performance of the proposed method in the fine-grained image stylization setting. Compared to the stacked multiple adapters approach, the proposed FGD module achieves better condition balance across all metrics. However, due to its static injection strategy, the quantitative results remain suboptimal. Other baselines show varying degrees of deterioration in image quality, style similarity, and text controllability. In contrast, **Co-Painter** achieves the best overall quantitative evaluation results, maintaining the optimal balance between SS (BSTS & COLS) and TA (CONS) while achieving the best IQ.

Qualitative Evaluation Analysis Figure 5 shows that **Co-Painter** outperforms all other baselines, achieving the best qualitative evaluation results. Composable adapters (T2I-Adapter [15] and ControlNet [33]) struggle with fine-grained color attributes, and conflicts between the canny map and text prompts cause significant discrepancies. Other baselines expose obvious issues, such as content leakage (IP-Adapter [32] (fine-tuning & stacked)) and style discrepancies (FGD, T2I-Adapter [15] (stacked), and DEADiff [19]). In contrast, **Co-Painter** captures brushstroke and color attributes from two reference images with high fidelity. Its fine-grained decoupling mechanism separates content, brushstroke, and color attributes, preventing information leakage. The gated feature injection mechanism adaptively incorporates multiple attributes, enhancing the capture of detailed information from reference images.

Difference Between FGD and Stacked Adapters It is noteworthy that the FGD module differs significantly from stacking multiple IP-Adapters [32] or T2I-Adapters [15]. The latter methods conflate the mapping of features from the CLIP [20] latent space to the attention space with the decoupling of style attributes, resulting in suboptimal decoupling or even overfitting. In contrast, the FGD module performs decoupling effectively by keeping the pre-trained image projection frozen to preserve feature space mapping

Method	Type	IQ↑	SS↑	TA↑	CONS↑	BSTS↑	COLS↑
ControlNet [33]	Controllable Model	5.76	19.5	22.2	20.5	19.3	18.0
T2I-Adapter [15]	Controllable Model	5.85	18.5	**25.7**	22.6	18.0	18.5
IP-Adapter [32]	Stylized Model	6.01	22.5	21.9	21.7	20.5	21.0
DEADiff [19]	Stylized Model	5.91	22.8	24.9	23.2	20.6	20.7
T2I-Adapter(stacked) [15]	Stylized Model	6.02	23.7	23.5	23.0	20.0	22.5
IP-Adapter(stacked) [32]	Stylized Model	6.03	**24.6**	18.2	18.9	**20.9**	**22.6**
CO-PAINTER(FGD)	Stylized Model	6.02	24.2	24.3	23.3	20.1	22.4
CO-PAINTER(Ours)	Stylized Model	**6.06**	24.4	24.5	**23.5**	20.6	**22.6**

Table 2. Quantitative evaluation results in a fine-grained setting. The image stylization results of controllable models are suboptimal. In contrast, stylized models demonstrate superior performance. Based on decoupling and adaptive injection, our method outperforms others across multiple metrics.

Method	IQ↑	SS↑	TA↑
Baseline	6.07	**36.9**	14.3
+GFI	5.96	31.6	23.3
+FGD	5.97	29.5	24.0
+FGD & GFI	**6.14**	27.8	**24.5**

Table 3. Evaluation results of ablation study. The combination of FGD & GFI achieves superior performance.

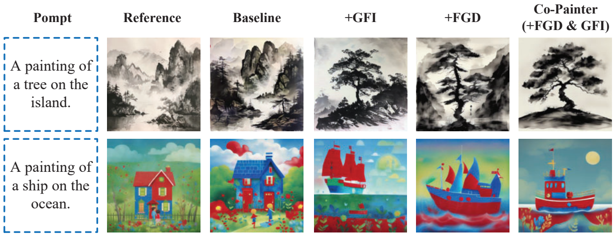

Pompt
Reference
Baseline
+GFI
+FGD
Co-Painter
(+FGD & GFI)
A painting of a tree on the island.
A painting of a ship on the ocean.

Figure 6. Qualitative evaluation of the ablation study. These results validate the importance of each component of **CO-PAINTER**.

while using three lightweight adapters focused on implicit fine-grained attribute decomposition.

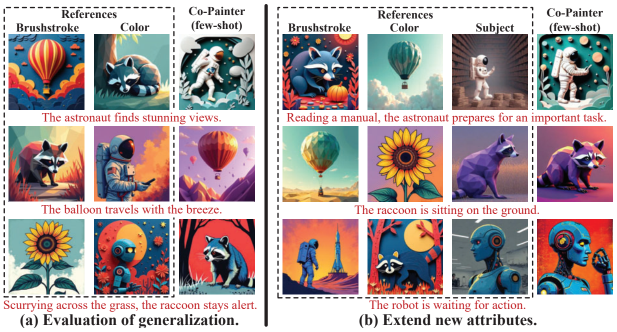
5.3. Ablation Study

We conduct a series of ablation studies to analyze the impact of each component in the **CO-PAINTER** on image stylization. Figure 6 and Table 3 present the results of the qualitative and quantitative evaluation, respectively.

Gated Feature Injection Mechanism. We first analyzed the impact of the gated feature injection mechanism on image stylization. It can be noted from Figure 6 (col 4) and Table 3 (row 2), that GFI utilizes a gated mechanism to adaptively integrate text and image features, effectively addressing the issue of image content leakage and condition imbalance. However, its performance remains limited due to the mixing of the reference image attributes.

Fine-Grained Decoupling Mechanism. We conducted a discussion of the FGD module. From Figure 6 (col 5) and Table 3 (row 3), it can be seen that fine-grained disentanglement of the image attributes further enhances the model's balance between text and image conditions. However, injecting the reference attribute information statically leads to the loss of detailed information (lower IQ) and slight content leakage (an unexpected mountain in Figure 6).

Combine FGD and GFI Mechanisms. From Figure 6 (col 6) and Table 3 (row 4), we can observe that the FGD module provided the model with a clearer, decoupled representation of image attributes. With this support, the GFI module can effectively capture detailed information about different conditions and their collaborative relationships. These two components are complementary in fine-grained image stylization tasks, and their integration yields the optimal image stylization performance.


References
Brushstroke
Color
Co-Painter
(few-shot)
The astronaut finds stunning views.
The balloon travels with the breeze.
Scurrying across the grass, the raccoon stays alert.
References
Brushstroke
Color
Subject
Co-Painter
(few-shot)
Reading a manual, the astronaut prepares for an important task.
The raccoon is sitting on the ground.
The robot is waiting for action.

(a) Evaluation of generalization.

(b) Extend new attributes.

Figure 7. Visualization results of **CO-PAINTER** evaluated on an out-of-domain few-shot dataset.

5.4. Out-of-Domain Generation

To evaluate the model's out-of-domain (OTD) performance, we performed few-shot fine-tuning on **CO-PAINTER** using OTD data (unseen contents, brushstrokes, and colors) and conducted model evaluation on the test set. There are only 10 training samples for each content type at different stylized levels (brushstroke, color). As Figure 7-(a) shows, our model exhibits exceptional generalization capability on OTD data. In addition, we also added a subject decoupling branch to validate the model's capability for attribute expansion. The evaluation results demonstrate that our lightweight structure enables easy extension to new attributes with just a linear and normalization layer (Figure 7-(b)). These parameters are negligible for the model.

6. Conclusion

In this paper, we introduce **CO-PAINTER**, an advanced image stylization model that achieves fine-grained control and enhanced flexibility through a unique decoupling mechanism and gated feature injection. Compared to previous works, the model was endowed with superior fine-grained controllability. Furthermore, it can also seamlessly integrate with other controllable models. By enabling precise control of the image style, **CO-PAINTER** paves the way for more sophisticated and personalized image generation.

Acknowledge This work was supported in part by the National Natural Science Foundation of China under Grant 62472359, Grant 62372379; in part by Xi'an's Key Industrial Chain Core Technology Breakthrough Project under Grant 24ZDCYJSGG0003.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024. 2

[3] Dar-Yen Chen, Hamish Tennent, and Ching-Wen Hsu. Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8619–8628, 2024. 1, 2, 6

[4] Bin Cheng, Zuhao Liu, Yunbo Peng, and Yue Lin. General image-to-image translation with one-shot image guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22736–22746, 2023. 2

[5] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11326–11336, 2022. 2, 6, 7

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1

[7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2

[8] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. 1

[9] Hexiang Hu, Kelvin CK Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, et al. Instruct-imagen: Image generation with multi-modal instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4754–4763, 2024. 2

[10] Nisha Huang, Fan Tang, Weiming Dong, and Changsheng Xu. Draw your art dream: Diverse digital art synthesis with multimodal guided diffusion. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1085–1094, 2022. 1, 2

[11] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. Diffusion model-based image editing: A survey. *arXiv preprint arXiv:2402.17525*, 2024. 1

[12] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6689–6700, 2024. 1

[13] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 1

[14] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 1

[15] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2, 6, 7, 8

[16] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 2

[17] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 2

[18] Fred Phillips and Brandy Mackintosh. Wiki art gallery, inc.: A case for critical thinking. *Issues in Accounting Education*, 26(3):593–608, 2011. 3

[19] Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8693–8702, 2024. 1, 2, 5, 6, 7, 8

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 6, 7

[21] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1, 2

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 6

[23] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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*, pages 22500–22510, 2023. 2

[24] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo

Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 3

[25] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3

[26] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8849, 2024. 1

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6

[28] Gemma Canet Tarrés, Dan Ruta, Tu Bui, and John Collomosse. Parasol: Parametric style control for diffusion image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2432–2442, 2024. 1

[29] Jagpal Singh Ubhi, Ashwani Kumar Aggarwal, et al. Neural style transfer for image within images and conditional gans for destylization. *Journal of Visual Communication and Image Representation*, 85:103483, 2022. 2

[30] Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 1, 2

[31] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7677–7689, 2023. 1

[32] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 5, 6, 7, 8

[33] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3, 4, 7, 8

[34] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–8, 2022. 6, 7

[35] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10156, 2023. 2, 6, 7

[36] Zhanjie Zhang, Quanwei Zhang, Wei Xing, Guangyuan Li, Lei Zhao, Jiakai Sun, Zehua Lan, Junsheng Luan, Yiling Huang, and Huaizhong Lin. Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7396–7404, 2024. 1, 2