Free-Lunch Color-Texture Disentanglement for Stylized Image Generation

Jiang Qin[1,*], Alexandra Gomez-Villa[3,4,*], Senmao Li[2,*,‡], Shiqi Yang[2,†], Yaxing Wang[2], Kai Wang[5,6,3,‡], Joost van de Weijer[3,4]

[1]Harbin Institute of Technology, China [2]VCIP, CS, Nankai University, China
[3]Computer Vision Center, Spain [4]Universitat Autònoma de Barcelona, Spain
[5]Program of Computer Science, City University of Hong Kong (Dongguan), China
[6]City University of Hong Kong, HK SAR, China
https://deepffff.github.io/sadis.github.io

Figure 1: Stylized images generated by our *training-free* method, *SADis*. **(Up)** Example of texture (left) and color (above) conditioned text-to-image (T2I) generation. This approach offers creators enhanced color control, including the use of color palettes as shown in the last two columns. **(Down)** Example with conditioning on content image (left) and color (above), showing it extends to color-only stylized image generation.

Abstract

Recent advances in Text-to-Image (T2I) diffusion models have transformed image generation, enabling significant progress in stylized generation using only a few style reference images. However, current diffusion-based methods struggle with *fine-grained* style customization due to challenges in controlling multiple style attributes, such as *color* and *texture*. This paper introduces the first tuning-free approach to achieve *free-lunch color-texture disentanglement* in stylized T2I generation, addressing the need for independently controlled style elements for the Disentangled Stylized Image Generation (*DisIG*) problem. Our approach leverages the *Image-Prompt Additivity* property in the CLIP image embedding space to develop techniques for separating and extracting Color-Texture Embeddings (*CTE*) from individual color and texture reference images. To ensure that the color palette of the generated image aligns closely with the color reference, we apply a whitening

[*]Equal contribution.
[†]Visiting researcher in Nankai University.
[‡]Corresponding authors: Senmao Li, Kai Wang

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

and coloring transformation to enhance color consistency. Additionally, to prevent texture loss due to the signal-leak bias inherent in diffusion training, we introduce a noise term that preserves textural fidelity during the Regularized Whitening and Coloring Transformation (*RegWCT*). Through these methods, our Style Attributes Disentanglement approach (*SADis*) delivers a more precise and customizable solution for stylized image generation. Experiments on images from the WikiArt and StyleDrop datasets demonstrate that, both qualitatively and quantitatively, *SADis* surpasses state-of-the-art stylization methods in the *DisIG* task.

1 Introduction

Stylized Image Generation [11, 30, 36] aims to transfer a style from a reference image to a target image. This field has evolved through several technological paradigms, beginning with CNN-based feature manipulation [20, 65, 36], advancing to attention mechanisms [4, 75, 46], and further developing through GAN-based image translation [37, 83, 73]. A significant advancement came with multimodal CLIP guidance [19, 40, 51], which bridged the gap between visual and textual representations. This breakthrough enabled a novel approach where style transfer could be guided by textual descriptions rather than being limited to reference images [19, 41, 35]. The field underwent another transformation with the emergence of Text-to-Image (T2I) diffusion models [56, 59, 57], which demonstrated remarkable capabilities in personalized image generation [58, 18, 48, 7].

T2I diffusion models catalyzed new developments in stylized image generation [54, 67, 62] — a specialized paradigm that focuses on creating new images that incorporate specific visual characteristics from reference styles, rather than simply transferring style between existing images. Initially, many approaches [18, 2] relied on extensive fine-tuning using datasets of similarly styled images. However, this requirement proved impractical in real-world scenarios, where collecting cohesive style-specific datasets is often challenging. Addressing these limitations, recent research has focused on developing *tuning-free (free-lunch)* methods [76, 54, 38]. These approaches eliminate the need for costly retraining while maintaining efficient style integration capabilities. Despite their methodological differences, both tuning-based and tuning-free approaches share a common characteristic: they transfer entangled style representations during the image generation process, simultaneously incorporating both color and texture elements from the style images into the final output.

Although existing stylized image generation methods offer promising advancements, for content creators, these methods have critical limitations. A key challenge is the lack of granular control over style elements. Content creators often need to adopt specific aspects of a reference style while preserving elements of their original vision [13, 27]. Color palettes, in particular, play a crucial role in this process — they are often meticulously crafted to evoke specific emotions or maintain brand consistency. Creators may wish to preserve these while adopting textural[4] elements from other reference styles. However, current approaches force an all-or-nothing choice, where accepting a reference style means incorporating all its visual (texture and color) characteristics simultaneously, seriously limiting artistic freedom and practical applicability.

To address these limitations, we introduce the problem of *Disentangled Stylized Image Generation* (*DisIG*), which aims to decompose reference styles into independently controllable attributes like color, texture, and semantic content. This formulation enables artists to selectively transfer specific style elements while maintaining control over the content using text prompts. One potential workaround for *DisIG* using existing methods is to specify desired style attributes through text prompts. However, this approach requires extensive prompt engineering [47, 77, 72] expertise and considerable trial-and-error to achieve results that can be obtained more intuitively through image prompts. Even with significant effort in crafting precise textual descriptions, text prompts often fall short of capturing the nuanced characteristics of the target style. This limitation stems from the inherent expressiveness gap between text and visual information — compared with text prompts, image prompts inherently contain more fine-grained semantic information that is difficult to describe

[4] In this paper, we use the term *texture* to refer to the arrangement, repetition, and local patterns of visual elements such as shapes, edges, intensities, and gradients and excluding color aspects.

accurately in text, including style attributes such as color and texture. Such nuanced details contribute to improved generation quality (as shown in Fig. 2-2nd row.).

In this work, we propose what we believe to be the *first DisIG* method that enables independent control of color and texture through separate reference images, without any training. Such application and performance can be seen from Fig. 1. To achieve this, we first analyze the *additivity property* — a characteristic previously studied in textual spaces [50, 6, 33] — and demonstrate that it also holds true in the *image prompt* space. Then, we work with each style component separately: we isolate the *color representation* through feature subtraction between the original color image and its grayscale equivalent, while deriving the *texture representation* using a grayscale version of the texture reference image and SVD rescaling. Finally, to ensure accurate color matching while preserving textures, we introduce *RegWCT*, an enhanced whitening and coloring transform with noise regularization, which counteracts the *signal-leak bias* [16, 80] in diffusion models.

Figure 2: Compared to SOTA stylization methods, *SADis* enables more precise and fine-grained control of color and texture by leveraging style element images, which allows for more accurate specification of visual attributes[5].

In experiments, we use images from the WikiArt [64] and StyleDrop [62] datasets and evaluate our approach based on the SDXL model [53]. By comparing against several state-of-the-art stylization methods, *SADis* consistently outperforms these baselines over both qualitative and quantitative results, particularly in achieving accurate color and texture expression for the stylized T2I generation. In summary, our contributions are as follows:

- We introduce the *Disentangled Stylized Image Generation* (*DisIG*) problem, a critical challenge in real-world applications that enables richer semantic control in diffusion-based stylization.
- We are the *first* to identify and use the *image-prompt additivity* property in the CLIP image encoder, showing how image features can be decomposed and combined similarly to text embeddings.
- We present the *first* tuning-free (free-lunch) *style attribute disentanglement* method, *SADis*, emphasizing color and texture as key style attributes. Our color-texture extraction (*CTE*) and regularized whitening-coloring transformation (*RegWCT*) techniques enable effective disentangled color-texture stylization without additional tuning.
- Through thorough qualitative and quantitative evaluations in standard style transfer benchmarks, *SADis* consistently outperforms existing baselines in realistic generation under the *DisIG* scenario.

2 Related Work

Stylized image generation [30, 11, 37, 83], also related to classical style transfer, aims to generate images with the artistic style of a reference image. Early research focused on statistical feature manipulation to transfer artistic styles from reference images. Gatys et al. [20] pioneered this direction by using covariance matrices for style representation. Following works like AdaIN [36] improved efficiency by transferring feature statistics between style and content feature maps, while WCT [45] introduced whitening and coloring transformations to match covariance matrices. Recent architectural innovations have enhanced stylization capabilities.

A significant paradigm shift occurred with the introduction of CLIP [55], enabling connections between text and image representations in a shared embedding space. This advancement spawned methods like ClipStyler [41] to combine global and patch-level CLIP losses for image stylization.

[5]Here we use the prompt "A ballerina" as an example. The prompt used in InstantStyle and DEADiff is a detailed description of the color reference image generated by GPT-4o [1].

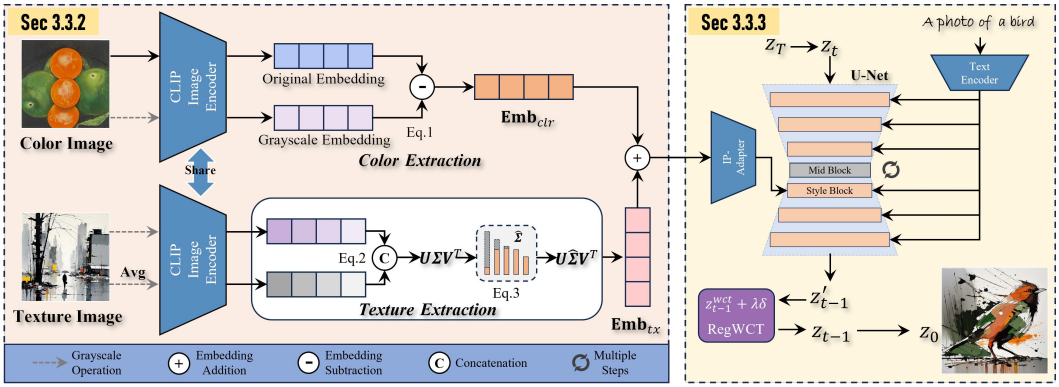

Figure 3: Our method, *SADis*, begins with color-texture extraction (*CTE*), leveraging the Image-Prompt Additivity property, which is verified in this paper for the *first* time. The color embedding $\mathbf{Emb}_{clr}$ is obtained by exploiting the Image-Prompt Additivity property, while the texture embedding $\mathbf{Emb}_{tx}$ is extracted via a singular value decomposition (SVD) operation. Afterwards, we incorporate these embeddings into the style cross-attention layer of the SDXL model. Subsequently, we refine the latent variable z_{t-1} at each inference step with our proposed *RegWCT* transformation, aligning color palettes precisely while retaining essential texture details.

Nowadays, T2I Diffusion models [61, 29, 8] have emerged as the new state-of-the-art models for text-to-image generation. Then the existing T2I stylization methods [23, 60, 12, 81, 66, 22, 82, 69, 43] achieve stylized image generation via fine-tuning generative models on few style reference images. However, this process is time-consuming and struggles to generalize to real-world scenarios where gathering a suitable subset of shared-style images can be difficult. To address these limitations, interest in tuning-free (free-lunch) approaches for stylized image generation has grown [10, 15, 70, 71, 26, 24]. These methods introduce lightweight adapters that extract style information from reference images and inject it into the diffusion process via self-attention or cross-attention layers. Representative examples include IP-Adapter [76] and Style-Adapter [71], which employ a decoupled cross-attention mechanism that separates cross-attention layers for handling text and image features independently. DEADiff [54] introduces a different approach by focusing on extracting disentangled representations of content and style using paired datasets.

Although existing stylization methods show promise, they lack granular control over individual style elements - a crucial need for content creators who often want to selectively adopt specific aspects of reference styles. This introduces a challenge we term the *Disentangled Stylized Image Generation* (*DisIG*). Focusing on color and texture as the most significant style attributes, this paper introduces the *first* free-lunch approach to *color-texture disentanglement* for stylized image generation.

3 Method

3.1 Preliminaries

T2I Diffusion Models. We built on the SDXL [53] model, consisting of two primary components: an autoencoder and a diffusion model $\epsilon_\theta(z_t, t, \tau_\xi(\mathcal{P}))$, where ϵ_θ is a UNet, conditioning a latent input z_t, a timestep $t \sim \mathrm{U}(1, T)$, and a text embedding $\tau_\xi(\mathcal{P})$. More specifically, text-guided diffusion models generate an image from the textual condition as $\mathcal{C}_{text} = \tau_\xi(\mathcal{P})$, where τ_ξ is the CLIP text encoder [55][6]. The cross-attention map is derived from $\epsilon_\theta(z_t, t, \mathcal{C}_{text})$. After predicting the noise, diffusion schedulers [63, 49] are used to predict the latent which we simplify as $z_{t-1} = \mathcal{G}(z_t, t, \mathcal{C}_{text})$.

IP-Adapter. Building on T2I diffusion models, the IP-Adapter [76] introduces additional controllability by conditioning the T2I model on a conditional image $\mathcal{I}_{ip}$. Practically, this involves leveraging a pre-trained T2I diffusion model and incorporating a cross-attention layer to the (projected) image condition following each text-prompt conditioning layer. The conditional image is encoded in the low-dimensional CLIP image embedding space [55] to capture high-level semantic information. By denoting the CLIP image encoder as τ_ϕ and IP-Adapter projection as $\mathbf{IP}$, this process is adding a new image condition $\mathcal{C}_{img} = \mathbf{IP}(\tau_\phi(\mathcal{I}_{ip}))$ to the T2I model as $z_{t-1} = \mathcal{G}(z_t, t, \mathcal{C}_{text}, \mathcal{C}_{img})$.

[6] SDXL uses two text encoders and concatenate the embeddings.

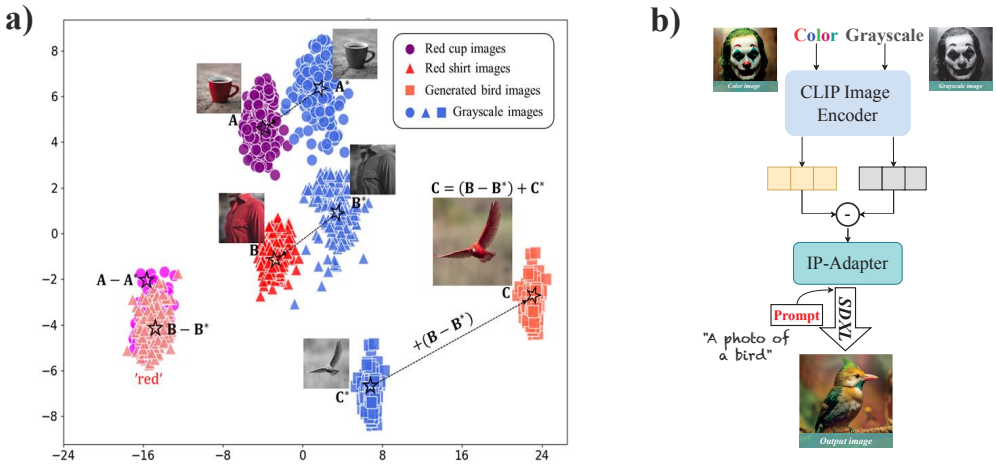

Figure 4: a) **Image-prompt additivity**: Image encoder latent space enables color extraction via grayscale subtraction. This color information can then be added to latent representations of other grey-scale images; b) By subtracting the grayscale image embeddings from the color image embeddings, we can effectively remove the texture information and isolate the color information.

3.2 Disentangled Stylized Image Generation

In a typical artistic workflow, creators draw inspiration from multiple sources to produce their final results. They might select a color palette from a sunset photograph, adopt textural patterns from a baroque painting, and incorporate structural elements from architectural blueprints—all while maintaining precise control over which elements they adopt from each reference. We term this ideal multi-target style transfer problem as Disentangled Stylized Image Generation (*DisIG*).

This multi-source approach to artistic creation (*DisIG*) contrast to current computational stylization methods [45, 9, 40], which typically extract and apply a bundled representation of "style" from a single reference image. To address the limitation, we first analyze how color and texture are entangled in existing stylization approaches. We focus specifically on these two attributes because they represent fundamental, measurable components of visual style that artists frequently manipulate independently.

Problem Setup. We formally define the *DisIG* task as follows: Given a text prompt $\mathcal{P}$, a color reference image $\mathcal{I}_{clr}$, and a texture reference image $\mathcal{I}_{tx}$, our goal is to generate an image that semantically aligns with $\mathcal{P}$ while independently adopting the color distribution of $\mathcal{I}_{clr}$ and the textural characteristics of $\mathcal{I}_{tx}$. The color reference $\mathcal{I}_{clr}$ can be either a standard RGB image or a discrete color palette. Unlike traditional style transfer that applies a holistic "style" from a single reference, our formulation enables precise, attribute-specific control by explicitly disentangling and separately transferring color and texture components.

3.3 Style Attributes Disentanglement

Here, we present our Style Attributes Disentanglement method (*SADis*). An overview is provided in Fig. 3. To successfully extract the color and texture information, we exploit the image-prompt additivity property as in Sec. 3.3.1. Next, we explain the color-texture extraction (*CTE*) in Sec. 3.3.2, and the noise-regularized whitening-coloring transformation (*RegWCT*) is detailed in Sec. 3.3.3.

3.3.1 Image-Prompt Additivity

To achieve disentangled stylization, we first analyze the *additivity property* of *image prompts*. This property is the main insight on which our color-texture disentanglement is based. It is illustrated using the CLIP image encoder, as adopted in the IP-Adapter [76]. Note that this property has been explored in text prompt spaces [50, 6]; here we show that it also holds in the image embedding space.

Consider the image embedding space as shown in Fig. 4-(a). Here we show a PCA plot of the embedding space for a distribution of images generated with three different text prompts. The embedding of a color image contains all information of the image (see e.g. points A and B). If we compare this embedding with the embedding of its grey-scale version (A', B'), the difference between the two embeddings is due to the removal of color information. We here hypothesize that

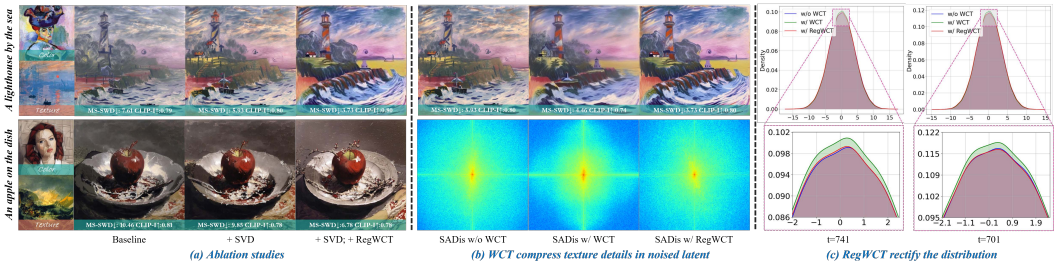

Figure 5: We extract texture information and suppress gray expression through SVD reweighting, and color alignment is improved by *RegWCT*.'Baseline' denotes color-texture disentanglement with Image-Prompt Additivity only (i.e., without SVD and *RegWCT*). (a) shows the Roles of SVD and *RegWCT* within *SADis*. (b) Our approach enables more precise color alignment with $\mathcal{I}_{clr}$ via *RegWCT*, whereas direct WCT manipulation compresses texture details and high-frequency components. (c) *RegWCT* rectifies the latent distribution that is affected by WCT. The distribution of the image latent generated with *RegWCT* technique is closer to that of the image generated without WCT in the frequency domain, across diverse time steps.

adding this difference to the embedding of another grey image, would transfer the color information to this image. This is illustrated for the figure of the grey bird image C. We can also see that the differences of $A - A'$ and $B - B'$ form a relatively compact distribution.

3.3.2 Color-Texture Extraction

Leveraging this image-prompt additivity, we disentangle color and texture representations from their respective reference images. For *color extraction*, we take the color image embedding and subtract the grayscale image embedding:

$$\mathbf{Emb}_{clr} = \tau_\phi(\mathcal{I}_{clr}) \ominus \tau_\phi(\mathbf{GS}(\mathcal{I}_{clr})) \tag{1}$$

where $\mathbf{GS}$ is the grayscale operation, $\mathbf{Emb}_{clr} \in \mathbb{R}^{n_t \times c}$. This strips away semantic information to retain only the color attributes (See Fig.4-b).

For *texture extraction*, we first convert the texture reference image to grayscale to remove any influence of its color palette, which is formulated as $\mathbf{Emb}^*_{tx} = \tau_\phi(\mathbf{GS}(\mathcal{I}_{tx}))$, $\mathbf{Emb}^*_{tx} \in \mathbb{R}^{n_t \times c}$. However, using only the grayscale texture embedding creates a mismatch in scale between the color and texture branches, often resulting in *overly gray tones* in T2I generations (as shown in Fig. 5 (a)-2nd col.). To address this, we average the gray texture image to obtain a pure gray image, which summarizes the main gray tone information in the $\mathbf{GS}(\mathcal{I}_{tx})$ image. Following that, we *concatenate* both embeddings from token-wise to obtain the initial texture representations:

$$\mathbf{Emb}^{'}_{tx} = \mathbf{Emb}^*_{tx} \copyright \tau_\phi(\mathbf{Avg}(\mathbf{GS}(\mathcal{I}_{tx}))) \tag{2}$$

$\mathbf{Emb}^{'}_{tx} \in \mathbb{R}^{2n_t \times c}$. Next, we conduct the Singular-Value Decomposition (SVD) over the texture representations. Inspired by [21, 44], we assume that the main singular values of $\mathbf{Emb}^{'}_{tx}$ correspond to the shared fundamental information of these two grayscale images, specifically the grayscale tone. We then have: $\mathbf{Emb}^{'}_{tx} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where $\boldsymbol{\Sigma} = diag(\sigma_0, \sigma_1, \cdots, \sigma_{n_t}, \cdots, \sigma_{2n_t-1})$,the singular values $\boldsymbol{\sigma}_0 \geq \cdots \geq \boldsymbol{\sigma}_{2n_t-1}$. To suppress the gray tone expression and extract the texture information, we introduce the augmentation for each singular value as:

$$\hat{\sigma} = \beta e^{-\gamma\sigma} * \sigma. \tag{3}$$

where e is the exponential, γ and β are parameters with positive numbers. We then recover it as $\mathbf{Emb}_{re} = \boldsymbol{U}\hat{\boldsymbol{\Sigma}}\boldsymbol{V}^T$, with the updated $\hat{\boldsymbol{\Sigma}} = diag(\hat{\sigma_0}, \hat{\sigma_1}, \cdots, \hat{\sigma}_{2n_t-1})$. The texture embedding is obtained by $\mathbf{Emb}_{tx} = \mathbf{Emb}_{re}[:n_t, :]$. That effectively reduces residual gray-color influences while preserving texture details. After extracting both color and texture reference image embeddings, we achieve a baseline of the *DisIG* problem. The T2I inference process is now formulated as: $z^{'}_{t-1} = \mathcal{G}(z_t, t, \mathcal{C}_{text}, \mathbf{Emb}_{tx} \oplus \mathbf{Emb}_{clr})$. Note that, we only inject the disentangled embedding $\mathbf{Emb}_{tx} \oplus \mathbf{Emb}_{clr}$ into *the first decoder layer* to compute cross-attention maps, which shows better *stylization* performance as previous works proved [67, 68, 2, 66]. Examples shown in Fig. 6 (2nd cols.) demonstrate the effectiveness of our proposal.

3.3.3 Regularized Whitening-Coloring Transforms

WCT: whitening-coloring transformation. Solely applying our *CTE* method does not fully ensure a precise color palette match between the reference color image $\mathcal{I}_{clr}$ and the generated output (see Fig.5). We hypothesize that while CLIP embeddings capture high-level semantic information, they may not fully represent nuanced color distribution details. To better align the color distributions of the generated image with the color reference, we incorporate a whitening-coloring transform (WCT) [45, 9, 31] applied to the noisy latent as $z_t^{wct} = \mathbf{WCT}(z_t^{'})$ (detailed WCT formulas in the Supplementary). This approach allows for a more faithful color transfer by aligning the latent representations with the reference color palette, enhancing the stylistic consistency of generated images. Examples of applying the WCT transform during the T2I generation process are presented in Fig. 5-(b), demonstrating an improved alignment of color palettes between the generated images and the reference color images. This enhancement visually confirms the effectiveness of WCT in achieving more accurate color transfer, resulting in a closer match to the desired color distribution.

Noise Regularization for WCT. However, as shown in Fig. 5-(b), the WCT process tends to compress texture details, distorting high-frequency information, which is also observed in previous works [80, 42]. This issue is attributed to the signal-leak bias [16] and the inverted latent distribution gap from the inversion process [80]. These terms essentially describe the same underlying phenomenon: during training, the distribution of z_T—the latent representation after adding Gaussian noise to z_0—does not perfectly match a standard Gaussian. Instead, it always contains the leakage or some certain prior information towards the original image, causing the residual structure to persist even at high noise levels. This biased noise in the forward diffusion process leads to a slight misalignment in texture representation during generation. To address this, we propose adding a small-scale noise to the latent as $z_t = z_t^{wct} + \lambda \cdot \delta, \delta \sim \mathcal{N}(0, 1)$, determined by a scale hyperparameter λ, to recover the lost high-frequency details. By introducing this noise regularization term, the regularized whitening-coloring transformation (*RegWCT*) achieves improved generative performance, better aligning both color and texture in the generated images with the reference images. Given a latent $z_t^{'}$ at timestep t, the color rectified latent z_t after *RegWCT* is formulated as $z_t = (1 - \omega)\, z_t^{'} + \omega \cdot \mathbf{RegWCT}(z_t^{'})$, where ω acts as a balance weight. To prevent the inference trajectory from deviating too much from the original one, we only perform the *RegWCT* transformation during the intermediate steps as $[T_{start}, T_{end}]$. With all these techniques, including color-texture extraction (*CTE*) and Regularized WCT transformation (*RegWCT*), our method *SADis* achieves customizable and flexible Disentangled Stylized Image Generation (as shown in Fig. 5 (a)-4th col.).

3.3.4 ControlNet-based real image stylization

Our method can be integrated with ControlNet [79] $\mathcal{CN}$ to facilitate content-based stylized image generation, as illustrated in Fig. 1 (down) and Fig. 7 (a). By using any pretrained ControlNet model (e.g. Canny-conditioned) as the base pipeline while maintaining all other hyperparameters consistent with *SADis*, we are able to significantly broaden the applicability of our method. More specifically, we have an input $\mathbf{I}_c$ as the content image, it is passed through the ControlNet $\mathcal{CN}(\mathbf{I}_c)$ as conditions for T2I generation model as $z_{t-1}^{'} = \mathcal{G}(z_t, t, \mathcal{CN}(\mathbf{I}_c), \mathbf{Emb}_{tx} \oplus \mathbf{Emb}_{clr})$, where $z_T \sim \mathcal{N}(0, 1)$ and the textual prompt as null $\mathcal{P} =$ "".

4 Experiments

4.1 Experimental Setups

Datasets. To ensure a fair comparison, we randomly select 40 images in total from the WikiArt [64] and StyleDrop [62] datasets. For our method, *SADis*, each of these images serves as either a color reference or a texture reference to enable color-texture disentanglement. In contrast, since the comparison methods lack disentanglement capabilities, we treat each of these 40 images as a style reference for these methods, supplemented by auxiliary text prompts generated from GPT-4o [1] as a strong captioning model. For quantitative comparisons, we used 20 images as the color reference set and 20 as the texture reference set. We also sampled 10 content prompts from StyleDrop, resulting in *4,000* stylized images per method to ensure fair and extensive comparisons with numerous images.

Figure 6: *Disentangled Stylized Image Generation* (*DisIG*) compared with baseline methods. For other approaches, we used GPT-4o to generate color descriptions based on the color reference image, incorporating these descriptions into the text prompts (details shown in the Supplementary). Our *SADis* accepts separate color and texture reference images to achieve flexible control.

Evaluation Metrics. We evaluate our method using multiple quantitative metrics. (1) For a general quality evaluation, we use CLIP-Score (CLIP) [28] to evaluate the T2I generation performance, specifically measuring the semantic consistency between the generated image and the text prompt. (2) For the evaluation of color attribute alignment, the MS-SWD [25] metric specifically evaluates color distance between generated and reference images. Color histogram distance (C-Hist) computes the distance between both color histograms. The GPT-4o color score [1] requests the GPT model to rate from 0-5 according to the color consistency between the color reference and the generated image. (3) For the texture quality, the Kernel-Inception Distance (KID) [5] is assessing the quality of generated images by measuring the dissimilarity between the real and generated image distributions (We use feature dimension as 2048 and evaluate on the 4000 images). CLIP-I is used to evaluate the similarity between the texture image and the generated image in grayscale.

Implementation Details. We build our method *SADis* upon the SDXL [53]. To apply the image embeddings as additional conditions to the T2I model, we use the IP-Adapter [76] pretrained projectors. Both of these models are based on the CLIP model [55], where the SDXL model utilizes the CLIP text encoder to generate textual embeddings and the IP-Adapter leverages the CLIP image encoder to extract image embeddings before the embedding projector. For the hyperparameters, we set $\lambda = 0.01, T_{start} = 0.8T, T_{end} = 0.6T, \omega = 0.5, \gamma = 0.003, \beta = 1.0$. For the time cost, we reported for 50-step DDIM. All the experiments are conducted on a single L40s GPU.

Comparison Methods. We compare with several state-of-the-art tuning-free stylization approaches based on the T2I diffusion models: DEADiff [54], InstantStyle [67], IP-Adapter [76], CSGO [74], DreamStyler [3], StyleDrop [62] and Artist [39]. For these baseline methods, we utilize the GPT-4o [1] model to generate captions from the color reference image, which are appended to the original textual prompt to guide the T2I generation. Additionally, SDXL [53] and Artist model are included in comparisons by generating captions from both the color and texture reference images, appending these captions to the textual prompts to guide the generation process. *We include much more experimental setup details, qualitative and quantitative results in the Supplmentary Material.*

4.2 Experimental Results

Qualitative Comparison. In our comparison of generation quality, visual results under the Disentangled Stylized Image Generation (*DisIG*) scenario are shown in Fig. 6 and Fig. 7. While baseline methods manage to capture some style information from the style reference image (the texture image for *SADis*), they fail to represent the color palette provided by the color reference accurately. The IP-Adapter and InstantStyle methods even retain original layouts and elements from the style images, deviating from the intended stylization. Other baselines, such as Artist [39] and DEADiff [54], exhibit poor generation quality, reflecting a limited understanding of both color and texture attributes. In contrast, our method, *SADis*, allows for both texture and color reference inputs, offering precise, con-

Table 1: Quantitative Comparison with existing image stylization methods. The best and second-best numbers are marked with **bold** and underlined respectively.

Method	CLIP	Color			Texture		Time Cost (s)	User study (%)		
		MS-SWD↓	C-Hist↓	GPT4o↑	CLIP-I↑	KID↓		Color↑	Texture↑	Both↑
SDXL [53]	0.272	9.51	1.20	3.01	0.69	0.08	9.26	16.99	5.84	11.65
IPAdapter [76]	0.233	11.54	1.23	2.84	**0.84**	**0.043**	9.52	6.08	22.54	14.55
InstantStyle [67]	0.261	12.53	1.32	2.80	0.74	0.056	9.31	4.81	24.94	13.10
Artist [39]	0.269	10.48	1.39	2.82	0.69	0.089	12.32	9.38	1.79	3.40
DEADiff [54]	0.267	11.20	1.24	2.73	0.69	0.087	**1.86**	4.57	2.59	2.67
StyleDrop [62]	0.275	13.52	1.43	2.67	0.70	0.054	6.91	5.07	3.57	5.34
DreamStyler [3]	0.277	12.17	1.26	2.39	0.71	0.060	5.23	4.67	3.57	5.39
CSGO [74]	0.280	14.25	1.36	2.63	0.69	0.071	15.99	6.59	9.73	6.06
SADis (Ours)	**0.281**	**5.57**	**0.96**	**3.34**	0.74	0.049	10.30	**41.83**	**25.42**	**37.84**

Table 2: Ablation by removing each components in our method *SADis*.

Method	Color		Texture	Time cost (s)
	MS-SWD↓	C-Hist↓	CLIP-I↑	
SADis (Ours)	**7.27**	**0.96**	0.74	≈ 10.30
− SVD	5.70	1.01	**0.76**	≈ 10.29
− *RegWCT*	8.05	1.06	0.75	≈ 9.42
− SVD − *RegWCT*	8.93	1.10	**0.76**	≈ 9.41

trollable T2I stylization. The generated images by *SADis* exhibit significantly improved color fidelity and texture detail compared to other approaches, demonstrating its effectiveness in disentangled style attribute transfer. To further validate *SADis*'s robustness, we conduct two tests: (1) swapping the color and texture reference images as shown in Fig. 7 (c), and (2) modifying their saturation and illumination levels, as shown in Fig. 7 (d). Unlike existing approaches such as InstantStyle, which cannot preserve image content under these conditions (detailed in *Supplementary*), *SADis* successfully applies color and texture independently while keeping the content largely unchanged.

Quantitative Comparison. A detailed quantitative comparison is provided in Tab. 1 to further support our findings. Our method, *SADis*, retains text-image alignment quality at a level comparable to the base SDXL model [53], as indicated by the CLIP score. This comparison demonstrates that *SADis* preserves alignment with the textual prompt better than other methods. Regarding color alignment, *SADis* significantly outperforms other approaches. For texture representation, the IP-Adapter-based methods, including InstantStyle [67] and *SADis*, show superior performance over other baselines. They achieve higher texture-related metrics at the expense of color fidelity and text-image alignment quality. The high texture scores for IP-Adapter can be explained by the fact that they control all cross-attention layers; However, this leads to high-semantic content leakage (like the same stars in the Van Gogh bear in Fig. 6). We decided to prevent this by only controlling the cross-attention of the low-level layers following InstantStyle [67, 2, 66]; Therefore, our texture results are close to those of InstantStyle, but at the same time we greatly improve color fidelity. In addition, their generated images follow the same structure as the original texture image, showing that they do not disentangle structure from texture but overfit to the contents of the texture images.

Ablation Study. The ablation study for each component of *SADis* is listed in Tab. 2. Results show that SVD rescaling in the *CTE* process and the noise-regularized *RegWCT* techniques significantly enhance color alignment while only slightly reducing texture precision, which is almost undetectable in the CLIP-I metric. The trade-off between texture and color alignment is optional for users, allowing them to adjust the balance based on the requirements.

User Study. To better assess alignment with human preferences, as shown in Tab. 1(right), we conducted a user study with 24 participants (30 sextuplets/user), collecting 720 data for each method. Each was asked to "select the best image from pairs of generated images, taking into account overall quality (considering *both* color and texture alignments), color alignment, and texture alignment, respectively." Our method outperformed other baseline approaches with at least a 20% improvement. This demonstrates the potency of *SADis* and its high alignment with human preference.

4.3 Additional Features

Image-based stylization. *SADis* Integrated with ControlNet achieves image-based stylized generation, as shown in Fig. 7 (a). Here, we adopt ControlNet (Canny) as the base model, setting its conditioning scale to 0.6. This integration broadens the application scenarios of *SADis*.

(a) Image based stylization (b) Color and material transfer (c) Swapping the color and texture (d) Saturation and Illuminant variation

Figure 7: (a) *SADis* allows for *real-image stylization*. (b) *SADis* effectively disentangles **color** and **material** elements from separate images, enabling precise control over color and material in image generation. (c) *SADis* remains effective when exchanging texture and color images. (d) *SADis* can capture color image properties such as saturation and illuminant. This further demonstrates the robustness of *SADis* in realistic *DisIG* scenario.

Prompt: A depiction of a woman dissolving into a flock of butterflies as she walks through a mystical forest

Prompt: A giant phoenix soaring across the sky, its feathers echoing the colors of the sunset

(a) FLUX.1-dev (b) SD3.5

Figure 8: SADis is compatible with MMDiT-based models. (a) demonstrates color-texture disentangled stylization results on FLUX.1-dev, while (b) shows the corresponding results on SD3.5.

Color and material transfer. Given color and material references from separate images, *SADis* is also capable of disentangling color and material elements, and provides precise controls over color and materials in T2I generation. The color and material transfer results are shown in Fig. 7 (b). This capability of *SADis* highlights its potential applications in artistic creation and industrial design.

Compatible with MMDiT-based models. *SADis* is also compatible with MMDiT-based models (SD3.5 [14] and FLUX.1-dev[7]) to achieve color-texture disentangled stylization. Specifically, we use SD3.5 and FLUX.1-dev as the base models. These models employ the original text encoders (i.e., T5, CLIP-L, and CLIP-G) for content control. We utilize SigLIP [78] with a pretrained adapter[8,9] for style control. Color-texture embeddings are extracted and disentangled from SigLIP features, following the same workflow as illustrated in Fig. 3. As shown in Fig. 8, SADis achieves color-texture disentanglement and enables stylized image generation on these MMDiT-based models as well. This demonstrates that the color-texture disentanglement capability of SADis does not rely on the U-Net architecture and can be generalized to MMDiT-based models for stylized image generation. It also indicates that SADis is not limited to specific CLIP models and can be applied to other text-image pretrained encoders.

5 Conclusion

We addressed a critical challenge in stylized image generation by introducing the concept of disentangled stylized image generation (*DisIG*). We focused on color and texture as key style attributes, presenting the first approach, named style attribute disentanglement (*SADis*), for independent control over these elements. Using the image-prompt additivity property, we proposed novel techniques, including the color-texture extraction (*CTE*) and regularized whitening-coloring transformation (*RegWCT*), to ensure enhanced color-texture consistency and more accurate results. Experimental evaluations demonstrate that *SADis* significantly outperforms existing stylization methods, both qualitatively and quantitatively. This work also opens new avenues for more flexible and customizable image generation, paving the way for future innovations for art creators.

[7]https://huggingface.co/black-forest-labs/FLUX.1-dev.
[8]https://huggingface.co/InstantX/SD3.5-Large-IP-Adapter.
[9]https://huggingface.co/InstantX/FLUX.1-dev-IP-Adapter.

Acknowledgements

We acknowledge project PID2022-143257NB-I00, financed by MCIN/AEI/10.13039/501100011033 and ERDF/EU, and the Generalitat de Catalunya CERCA Program, and ELLIOT project funded by the European Union under Grant Agreement 101214398. This work was also supported by NSFC (NO. 62225604) and Youth Foundation (62202243). We acknowledge "Science and Technology Yongjiang 2035" key technology breakthrough plan project and Chinese government-guided local science and technology development fund projects (scientific and technological achievement transfer and transformation projects) (254Z0102G). Kai Wang acknowledges the funding from Guangdong and Hong Kong Universities 1+1+1 Joint Research Collaboration Scheme and the start-up grant B01040000108 from CityU-DG.

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.

[2] Aishwarya Agarwal, Srikrishna Karanam, Tripti Shukla, and Balaji Vasan Srinivasan. An image is worth multiple words: Multi-attribute inversion for constrained text-to-image synthesis. *International Conference on Machine Learning*, 2024.

[3] Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong. Dreamstyler: Paint by style inversion with text-to-image diffusion models. *Proceedings of the Conference on Artificial Intelligence*, 2024.

[4] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024.

[5] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

[6] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 25365–25389. Curran Associates, Inc., 2023.

[7] Muhammad Atif Butt, Kai Wang, Javier Vazquez-Corral, and Joost van de Weijer. Colorpeel: Color prompt learning with diffusion models via color and shape disentanglement. In *European Conference on Computer Vision*, pages 456–472. Springer, 2025.

[8] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

[9] Wonwoong Cho, Sungha Choi, David Keetae Park, Inkyu Shin, and Jaegul Choo. Image-to-image translation via group-wise deep whitening-and-coloring transformation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10639–10647, 2019.

[10] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8795–8805, 2024.

[11] 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.

[12] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022.

[13] Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. Art and the science of generative ai. *Science*, 380(6650):1110–1111, 2023.

[14] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[15] Martin Nicolas Everaert, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, and Radhakrishna Achanta. Diffusion in Style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2251–2261, October 2023.

[16] Martin Nicolas Everaert, Athanasios Fitsios, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, and Radhakrishna Achanta. Exploiting the signal-leak bias in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4025–4034, 2024.

[17] 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, 2022.

[18] 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. *International Conference on Learning Representations*, 2023.

[19] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022.

[20] 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.

[21] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2869, 2014.

[22] Inhwa Han, Serin Yang, Taesung Kwon, and Jong Chul Ye. Highly personalized text embedding for image manipulation by stable diffusion. *arXiv preprint arXiv:2303.08767*, 2023.

[23] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *Proceedings of the International Conference on Computer Vision*, 2023.

[24] Feihong He, Gang Li, Mengyuan Zhang, Leilei Yan, Lingyu Si, Fanzhang Li, and Li Shen. Freestyle: Free lunch for text-guided style transfer using diffusion models. *arXiv preprint arXiv:2401.15636*, 2024.

[25] Jiaqi He, Zhihua Wang, Leon Wang, Tsein-I Liu, Yuming Fang, Qilin Sun, and Kede Ma. Multiscale sliced wasserstein distances as perceptual color difference measures. *European Conference on Computer Vision*, 2024.

[26] 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 (CVPR)*, pages 4775–4785, June 2024.

[27] Aaron Hertzmann. Toward modeling creative processes for algorithmic painting. *arXiv preprint arXiv:2205.01605*, 2022.

[28] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021.

[29] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[30] Kibeom Hong, Seogkyu Jeon, Junsoo Lee, Namhyuk Ahn, Kunhee Kim, Pilhyeon Lee, Daesik Kim, Youngjung Uh, and Hyeran Byun. Aespa-net: Aesthetic pattern-aware style transfer networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22758–22767, 2023.

[31] Maliha Hossain. Whitening and coloring transformations for multivariate gaussian data. *A slecture partly based on the ECE662 Spring*, 2014.

[32] Miliha Hossain. Whitening and coloring transforms for multivariate gaussian random variables. *Project Rhea*, 3:1–7, 2016.

[33] Taihang Hu, Linxuan Li, Joost van de Weijer, Hongcheng Gao, Fahad Shahbaz Khan, Jian Yang, Mingming Cheng, Kai Wang, and Yaxing Wang. Token merging for training-free semantic binding in text-to-image synthesis. In *Advances in Neural Information Processing Systems*, 2024.

[34] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 2023.

[35] Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Weiming Dong, and Changsheng Xu. Diffstyler: Controllable dual diffusion for text-driven image stylization. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[36] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[37] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[38] Jaeseok Jeong, Junho Kim, Yunjey Choi, Gayoung Lee, and Youngjung Uh. Visual style prompting with swapping self-attention. *arXiv preprint arXiv:2402.12974*, 2024.

[39] Ruixiang Jiang and Changwen Chen. Artist: Aesthetically controllable text-driven stylization without training. *arXiv preprint arXiv:2407.15842*, 2024.

[40] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided image manipulation using diffusion models. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[41] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[42] Senmao Li, Taihang Hu, Fahad Shahbaz Khan, Linxuan Li, Shiqi Yang, Yaxing Wang, Ming-Ming Cheng, and Jian Yang. Faster diffusion: Rethinking the role of unet encoder in diffusion models, 2023.

[43] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Stylediffusion: Prompt-embedding inversion for text-based editing. *arXiv preprint arXiv:2303.15649*, 2023.

[44] Senmao Li, Joost van de Weijer, Fahad Khan, Qibin Hou, Yaxing Wang, et al. Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

[45] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017.

[46] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6649–6658, 2021.

[47] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–23, 2022.

[48] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *International Conference on Machine Learning*, 2023.

[49] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

[50] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[51] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *Proceedings of the International Conference on Computer Vision*, 2021.

[52] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[53] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[54] 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.

[55] 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.

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

[57] 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 (CVPR)*, pages 10684–10695, 06 2022.

[58] 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. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[59] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022.

[60] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[61] Alex Shonenkov, Misha Konstantinov, Daria Bakshandaeva, Christoph Schuhmann, Ksenia Ivanova, and Nadiia Klokova. Deepfloyd-if. `https://github.com/deep-floyd/IF`, 2023.

[62] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *Advances in Neural Information Processing Systems*, 2023.

[63] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[64] Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1):394–409, 2019.

[65] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022.

[66] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023.

[67] Haofan Wang, Matteo Spinelli, 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.

[68] Haofan Wang, Peng Xing, Renyuan Huang, Hao Ai, Qixun Wang, and Xu Bai. Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. *arXiv preprint arXiv:2407.00788*, 2024.

[69] Kai Wang, Fei Yang, Shiqi Yang, Muhammad Atif Butt, and Joost van de Weijer. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *Advances in Neural Information Processing Systems*, 2023.

[70] 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.

[71] Zhouxia Wang, Xintao Wang, Liangbin Xie, Zhongang Qi, Ying Shan, Wenping Wang, and Ping Luo. Styleadapter: A single-pass lora-free model for stylized image generation. *International Journal of Computer Vision*, 2023.

[72] Sam Witteveen and Martin Andrews. Investigating prompt engineering in diffusion models. *arXiv preprint arXiv:2211.15462*, 2022.

[73] Zongze Wu, Yotam Nitzan, Eli Shechtman, and Dani Lischinski. Stylealign: Analysis and applications of aligned stylegan models, 2021.

[74] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *arXiv 2408.16766*, 2024.

[75] Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang. Attention-aware multi-stroke style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1467–1475, 2019.

[76] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *Proceedings of the Conference on Artificial Intelligence*, 2024.

[77] Chang Yu, Junran Peng, Xiangyu Zhu, Zhaoxiang Zhang, Qi Tian, and Zhen Lei. Seek for incantations: Towards accurate text-to-image diffusion synthesis through prompt engineering. *arXiv preprint arXiv:2401.06345*, 2024.

[78] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[79] 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.

[80] Yuechen Zhang, Jinbo Xing, Eric Lo, and Jiaya Jia. Real-world image variation by aligning diffusion inversion chain. *Advances in Neural Information Processing Systems*, 36, 2024.

[81] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Expanded conditioning for the personalization of attribute-aware image generation. *SIGGRAPH Asia 2023*, 2023.

[82] 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.

[83] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Abstract and Introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix A.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Sec.3.2

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sec. 4.1

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Supplementary Material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 4.1

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Not applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix (Supplementary Material).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully checked the NeurIPS code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix A.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Not applicable.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We politely cited the existing assets and read their usage license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Not applicable.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Sec.4.2 and Supplementary Material.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Supplementary Material.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

A Statements

Limitations. The present study focuses on disentangling the color and texture elements in a training-free manner, and performing stylized image generation using these elements. This method can offer a flexible and customizable image generation for art creators and designers. However, when applying to content-consistency stylization, the consistency generation ability of this work can be further improved. In our future work, we will explore deeper with the consistency and expand our research into content-consistency stylization.

Broader Impacts. *SADis* enhances the flexible stylization capability in text-to-image synthesis by disentangling the color and texture elements. However, it also carries potential negative implications. It could be used to generate false or misleading images, thereby spreading misinformation. If *SADis* is applied to generate images of public figures, it poses a risk of infringing on personal privacy. Additionally, the automatically generated images may also touch upon copyright and intellectual property issues.

Ethical Statement. We acknowledge the potential ethical implications of deploying generative models, including issues related to privacy, data misuse, and the propagation of biases. All models used in this paper are publicly available. We will release the modified codes to reproduce the results of this paper. We also want to point out the potential role of customization approaches in the generation of fake news, and we encourage and support responsible usage.

Reproducibility Statement. To facilitate reproducibility, we will make the entire source code and scripts needed to replicate all results presented in this paper available after the peer review period. We will release the code for the novel color metric we have introduced. We conducted all experiments using publicly accessible datasets. Elaborate details of all experiments have been provided in the Appendices.

B Image-Prompt Additivity

B.1 Broader Property of Image-Prompt Additivity

We demonstrate the broader property of Image-Prompt Additivity in Fig. 9. For instance, in Fig. 9-(a), subtracting the embedding of a hat from the embedding of a person wearing a hat results in the generation of a person without a hat. Similarly, in Fig. 9-(b), adding the embedding of glasses to a person results in the generation of that person wearing glasses. We refer to this phenomenon as Image-Prompt Additivity. We hypothesize this property originates from the image-text paired training in CLIP models [55]. The training process endows the image branch with the additivity property inherent to the text embedding space [50, 33, 6]. Although the person identities are altered after embedding additivity manipulations, this approach demonstrates a promising property for enabling our training-free color-texture disentanglement. Building on this property, we develop our method, *SADis*, to extract color and texture information from reference images and effectively apply these attributes in T2I generation.

In future work, we aim to address the limitations of Image-Prompt Additivity by enhancing identity consistency after additivity manipulations. This could broaden the impact and applicability of this property, advancing its utility in the field of image generation.

B.2 Extra Analysis on Image-Prompt Additivity

As another illustration of image prompt additivity, we construct a set of blue, red, and green images. For example, the blue set is constructed by setting the red and green color channels to zero of a set of 100 images (similarly, for the green and red set). We isolate the *color representation* by performing a feature subtraction between the blue image and its grayscale equivalent. We plot their projected embeddings as blue dots in Fig. 10-(Right); note how these embeddings always maintain close to the pure blue embedding and are far from the green and red embeddings. Also, generated images with these subtraction blue color embeddings keep the bluish color palette as expected (see the car and t-shirt). An extended analysis is presented in Fig. 11, where we apply similar manipulations using

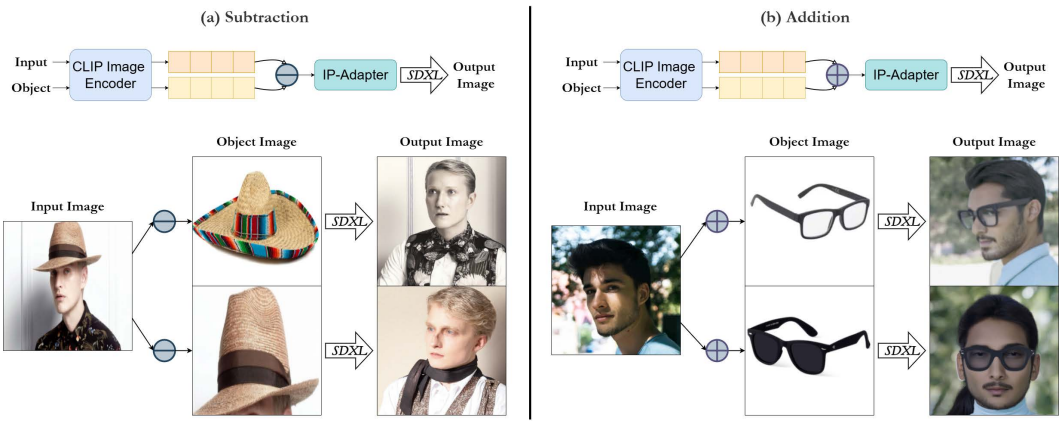

Figure 9: By subtracting or adding the object image embeddings from the input image embeddings, we can effectively remove or add the object to the scene, although some degree of identity information for the person is also diminished.

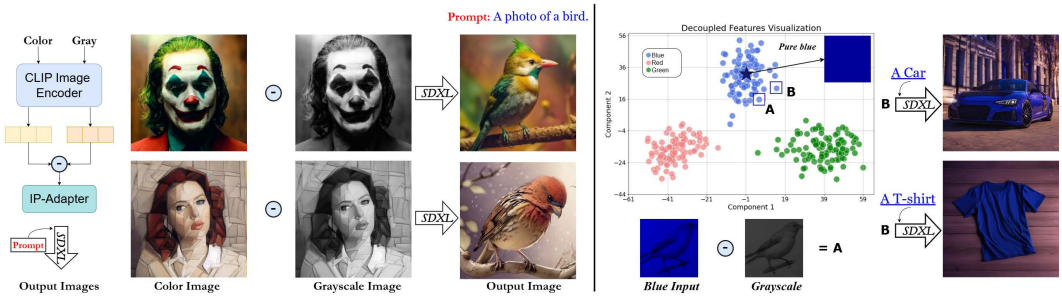

Figure 10: (Left) By subtracting the grayscale image embeddings from the color image embeddings, we can effectively remove the texture information and isolate the color information. Combined with the prompt "A photo of a bird" we can generate images in the same color schemes as the input images. (Right) By subtracting the bluish image embeddings with the grayscale embeddings for random generated images with 100 prompts, we visualize the subtracted embedding via PCA decomposition. We observe them gathering around the pure blue image. The generation with these subtracted embeddings further prove the consistency.

colorful images as color references. The successful clustering further confirms that the resulting embeddings effectively preserve essential color information.

C Implementation details

We develop our method, *SADis*, based on the SDXL model [53], which is among the leading open-source T2I generative models available. To construct the experimental datasets, we randomly select 40 images from the WikiArt dataset [64] and StyleDrop [62] image collections. Each image can be used either as a color reference or a texture reference in the experiments. All input images are resized to 512×512 before feeding into models. Also, for quantitative comparisons, 20 images are designated as the color reference set, while the remaining 20 serve as the texture reference set. Additionally, we sample 10 content prompts from StyleDrop, yielding a total of 4000 stylized images for each method for comparison. Since some comparison methods (such as SDXL [53] and Artist [39]) require text prompts as style controls for image generation, we employ the state-of-the-art vision-language model GPT-4o as the image captioning tool to generate precise color and texture prompts from the reference images. Subsequently, the content, color, and texture prompts are concatenated in the format: *'{content prompt}, {texture prompt} in the color {color prompt}'*. As for the comparison models (like DEADiff [54], IP-Adapter [76], Instantstyle [67], DreamStyler [3], StyleDrop [62], and CSGO [74]) that require text prompts for color control, the text prompt is constructed as: *'{content prompt}, in the color {color prompt}'*. All experiments were conducted on an NVIDIA-L40s GPU.

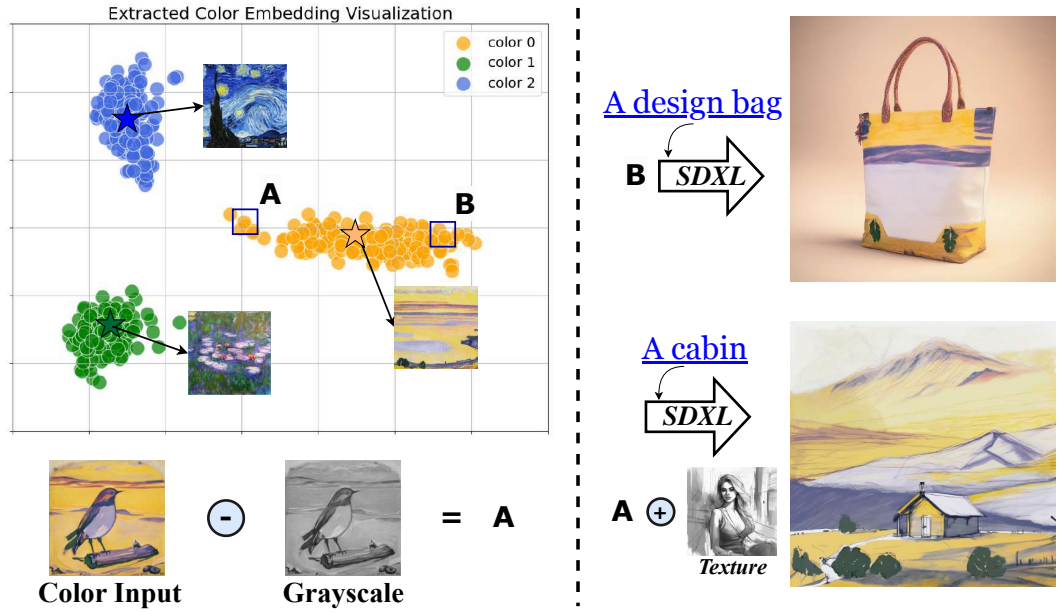

Figure 11: By subtracting the color image embeddings with the grayscale embeddings for random generated images with 100 prompts, we visualize the subtracted embedding via PCA decomposition. We observe them gathering around the color reference image. The generation with these subtracted embeddings further prove the consistency.

We further include the details of each evaluation metric. **CLIP Score** [55] is used to evaluate the semantic alignment between the text prompt and the generated image. We calucate the CLIP Score according to configuration of the T2I-CompBench repository [34]. For the color alignment evaluation, we use MS-SWD [25], color histogram distance (C-Hist), and GPT4o to calculate the color attribute similarity between the color reference and the generated image. **MS-SWD** [25]: since the generated imgae are usually not spatially aligned with the color reference image, we use MS-SWD to better evaluate the color attribute alignment according to the default setting of their repository [25]. **C-Hist:** we first compute the RGB histograms of the color reference and generated images. Afterwards, the Bhattacharyya distance is used to measure the differences between their color histograms. **GPT-4o score** [1]: to comprehensively evaluate the color alignment performance, we furture adopt the multimodal model GPT-4o to compute the color alignment score between the generated image and the color reference image. Specifically, the GPT-4o metric is computed according to Fig. 12. Firstly, following the previous work[17], we extract dominant colors by ColorThief. Afterwards, we feed the extracted color names to GPT-4o as the reference color from the color image, and ask GPT-4o to give 1-5 points according to the criteria shown in Fig. 12.

D WCT transformations formulas

Given a latent $z_t^{'} \in \mathcal{R}^{C\times H\times W}$ and a color reference latent $z_t^c \in \mathcal{R}^{C\times H\times W}$ at timestep t, we adopt Whitening and Coloring Transforms (WCT) [45, 9, 31] to transform $z_t^{'}$ to match covariance matrix of color reference latent z_t^c. There are two step for WCT: **Whitening transform** and **Coloring transform**.

Whitening transform: the latent $z_t^{'}$ is firstly centered by subtracting its mean vector m, and then an uncorrelated latent $\hat{z}_t^{'}$ is obtained by:

$$\hat{z}_t^{'} = ED^{-1/2}E^{\mathrm{T}}z_t^{'}, \tag{4}$$

where D denotes a diagonal matrix of $z_t^{'}z_t^{'\mathrm{T}} \in \mathcal{R}^{C\times C}$ and E is the corresponding orthogonal matrix of eigenvectors which satisfy $z_t^{'}z_t^{'\mathrm{T}} = EDE^{\mathrm{T}}$.

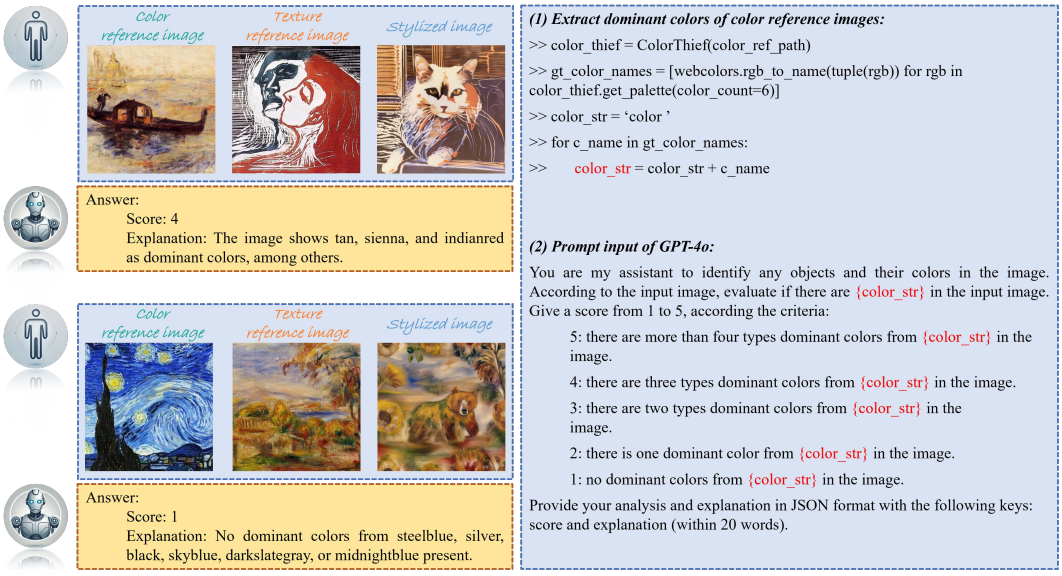

Figure 12: GPT-4o metric for color alignment evaluation.

Figure 13: Some color and texture prompts generated by GPT-4o for comparison methods in the main paper for qualitative evaluation.

Coloring transform: it's the reverse process of Whitening transform [32, 45]. Beforehand, the color reference latent z_t^c is centered by subtracting its mean vector m_c. Subsequently, we obtain the transformed latent z_t^{wct} which satisfies the desired correlations $z_t^{wct} z_t^{wct\mathrm{T}} = z_t^c z_t^{c\mathrm{T}} = I$:

$$z_t^{wct} = E_c D_c^{-1/2} E_c^{\mathrm{T}} \hat{z}_t^{'}, \tag{5}$$

where D_c is a diagonal matrix with the eigenvalues of the covariance matrix $z_t^c z_t^{c\mathrm{T}}$ and E_c is the corresponding orthogonal matrix of eigenvectors. Finally, we re-center the WCT transformed latent z_t^{wct} by adding the mean vector m_c of the color reference latent z_t^c:

$$z_t^{wct} = z_t^{wct} + m_c. \tag{6}$$

E Additional Experimental Results

In Fig. 14, we include more color-texture disentanglement examples of our method *SADis*, which is under the Disentangled Stylized Image Generation (*DisIG*) scenario.

Additional comparison results with complex prompts. To further evaluate performance with complex prompts, we conducted additional experiments by randomly sampling 10 complex prompts from DREAMBENCH++ [52], together with 10 color and 10 texture images from our dataset, resulting in 1000 generated images for evaluation. As shown in Tab. 3, compared to other methods,

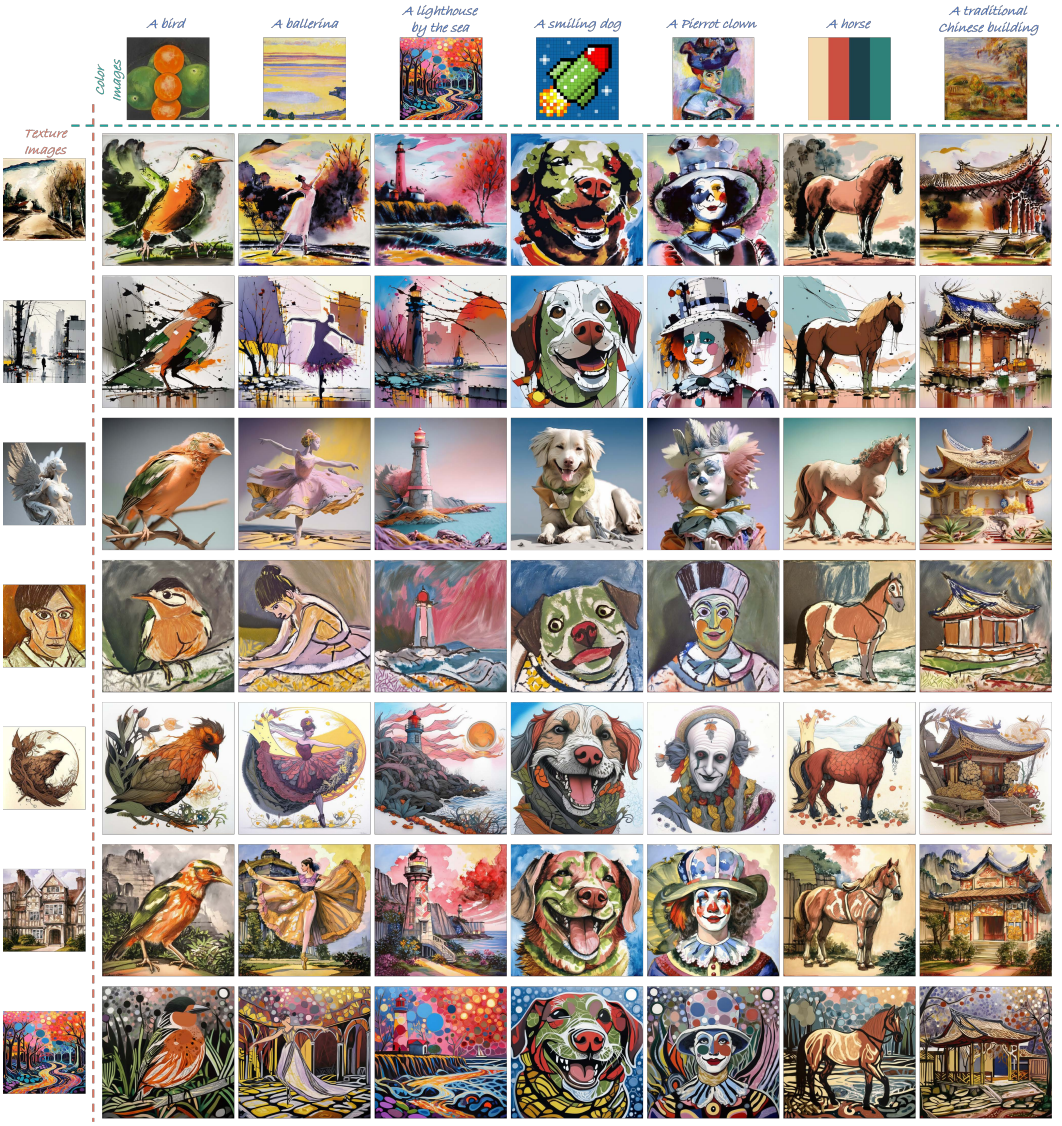

Figure 14: Additional experimental results of *SADis*.

our approach achieves the best disentanglement of color and texture, resulting in a more balanced performance in both color and texture consistency. Specifically, our method attains significantly higher color consistency, while other methods often exhibit strong color-texture entanglement, with color being overly influenced by the texture reference image. Although IP-Adapter achieves the highest numerical score for texture consistency, it suffers from severe semantic leakage from the texture reference image (Row 2 and Row 3 in Fig. 6), leading to poor text-image alignment (CLIP score: 0.257), which is inadequate for stylization. In contrast, except for IP-Adapter, our method achieves the highest scores in both texture consistency and text-image alignment.

Image-based stylization. Furthermore, our method, *SADis*, can be seamlessly integrated with ControlNet [79] to enable image-based stylized generation and material transfer, as demonstrated in Fig. 16 and Fig. 15(the last row) respectively. Here, we adopt ControlNet (Canny) as the base pipeline, setting its conditioning scale to 0.6. All other hyper-parameters are kept consistent with those of *SADis*. This integration broadens the application scenarios of our proposed approach.

Color and material transfer. Given color and material references from separate images, our method, *SADis*, is also capable of disentangling color and material elements, and provides precise controls over color and materials in T2I generation. The color and material transfer results are shown in

Figure 15: Stylized images generated by our *training-free* method, *SADis*. (Up) As shown in the first rows, it enables disentangled control over *color* and *style* attributes in text-to-image diffusion models using separate image prompts. This approach offers creators enhanced color control, including the use of color palettes as in the last two columns. (Down) *SADis* also enables real-image stylization by incorporating a content image as an additional condition via ControlNet. Furthermore, it extends to color-only stylized generation and material transfer for more flexible image generation.

Table 3: Quantitative comparison with complex prompts. The complex prompts are sampled from DREAMBENCH++ [52]. The best and second-best numbers are marked with **bold** and underlined respectively.

Method	CLIP	Color			Texture	
		SWD↓	C-Hist↓	GPT4o↑	CLIP-I↑	KID↓
SDXL [53]	0.291	9.00	1.14	3.07	0.698	0.091
IPAdapter [76]	0.257	9.99	1.15	2.96	**0.817**	**0.058**
InstantStyle [67]	0.278	11.21	1.25	2.85	0.751	0.065
Artist [39]	0.253	11.15	1.24	2.67	0.742	0.082
DEADiff [54]	0.280	10.01	1.14	2.80	0.718	0.090
StyleDrop [62]	0.290	11.90	1.35	2.83	0.731	0.078
DreamStyler [3]	0.300	11.27	1.21	2.57	0.677	0.083
CSGO [74]	0.299	12.60	1.22	2.64	0.728	0.081
SADis (Ours)	**0.301**	**6.08**	**0.95**	**3.14**	0.751	0.064

Fig. 19 and Fig. 15(the last row). This capability of *SADis* highlights its potential applications in artistic creation and industrial design. We also desire to note that, the term "color palette" typically refers to the overall color scheme of an image, not specific local regions. Our *SADis* does not control the color palette of individual objects in general. However, Fig. 19 demonstrates controlled color and material transfer using a masking mechanism with our method *SADis*.

Color and texture transfer from the same image. *SADis* performs exceptionally well when using the same image as both the color and texture reference (as shown in Fig. 20), showcasing its remarkable flexibility and adaptability. This ability highlights the method's robustness in leveraging a single source to effectively guide both color and texture information, ensuring consistent and coherent results. The comparative experimental results are presented in Tab. 4. In terms of texture consistency, our method achieves the best performance among all methods except for IP-Adapter. However, it is important to note that although the IP-Adapter achieves the highest numerical score for texture consistency, it introduces severe semantic leakage from the texture reference image (see Row. 2 and Row. 3 of IP-adapter results in Fig. 6), resulting in poor text-image alignment (CLIP: 0.260 in Tab. 4), which fails to meet the requirements of stylization. In contrast, excluding the IP-Adapter, our method achieves the highest scores in color consistency (MS-SWD: 3.19), texture consistency (CLIP-I: 0.754), and text-image alignment (CLIP: 0.302).

Robust to color variations. As shown in Fig.21 and Fig. 22, we vary the saturation and illuminant continuously. To be specific, for saturation, we developed a Python-based saturation adjustment tool

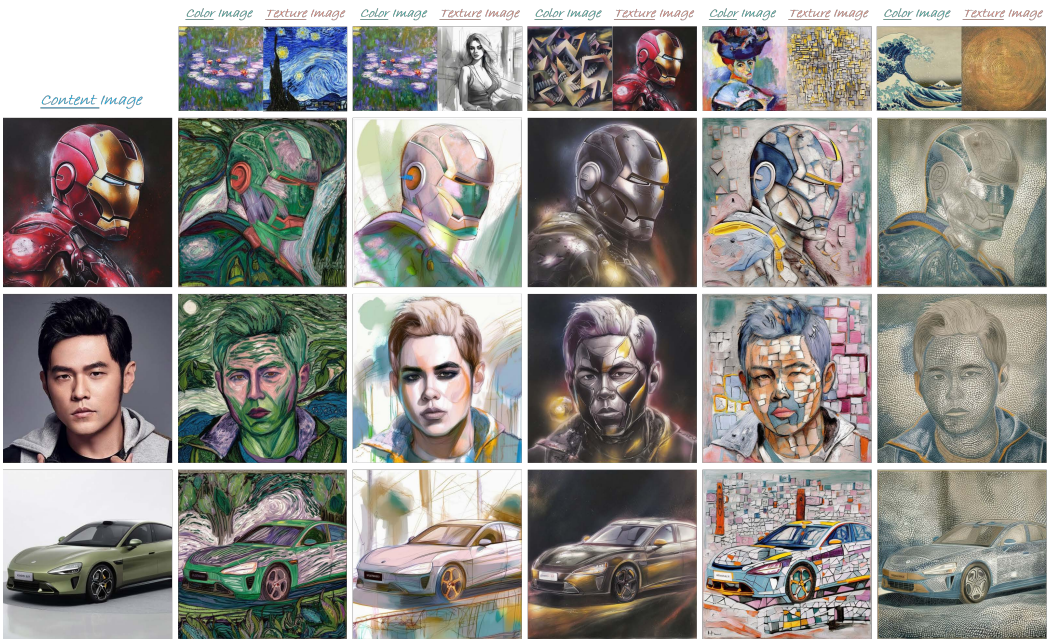

Figure 16: Our work is compatible with ControlNet to achieve **image-based stylization**.

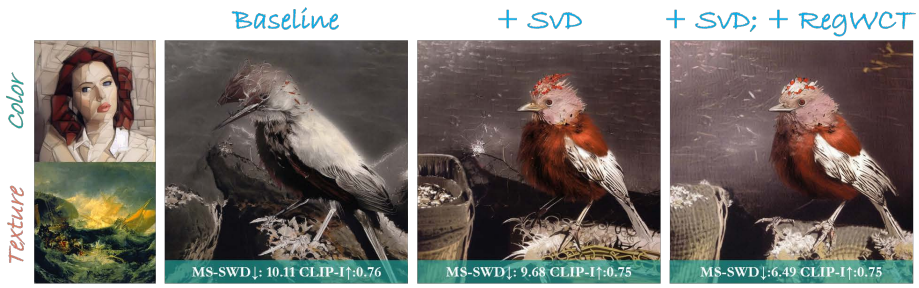

Figure 17: Visualization of ablating each component of *SADis*.

utilizing HSV (Hue, Saturation, Value) color space transformation. The tool linearly modifies the saturation channel using factors of [0.2, 0.6, 1.0, 1.5, 3.0], respectively, before converting the images back to RGB format. The results of saturation variations are shown in Fig. 21. Regarding the color of the illuminant, we modify images by applying calibrated RGB channel multipliers. For warm temperatures (red) as shown in Fig. 22, red increases by 30%, green reduces by 10%, and blue by 20%. The intensity of these adjustments can be controlled (0-1 range), enabling fine-grained color temperature manipulation. We present 3 increasing intensities in Fig. 22. Our method *SADis* is able to generate images with smooth change along with the saturation and illuminant color variations, while the other method InstantStyle [67] fails. That further proves the robustness of our method *SADis* for Disentangled Stylized Image Generation (*DisIG*).

Each component works independently. The T2I generations in Fig. 18-left provide evidence that the color and texture branches in *SADis* function independently. Additionally, the visualization for our ablation study in Fig. 17 further demonstrates improved color alignment with minimal texture degradation, achieving a good trade-off. However, the trade-off between texture and color alignment is optional for users, allowing them to adjust the balance based on the requirements of specific application scenarios.

Stability of different sampling rounds. We conducted additional experiments to assess the stability of color and texture in repeated generations. For each setting, the process was repeated five times, generating 1,000 images per round and computing the relevant metrics. As summarized in Tab. 5, the results show minimal variation in color and texture consistency across rounds, demonstrating the robustness and stability of our method.

Figure 18: *SADis* can control color and texture generation separately using reference images.

Figure 19: *SADis* effectively disentangles **color** and **material** elements from separate images, enabling precise control over color and material in image generation.

Figure 20: *SADis* performs well using the same image as both the color and texture reference, demonstrating its flexible capability.

Table 4: Quantitative comparison where both color and texture are derived from the same image. The best and second-best numbers are marked with **bold** and underlined, respectively.

Method	CLIP	Color			Texture	
		MS-SWD↓	C-Hist↓	GPT4o↑	CLIP-I↑	KID↓
SDXL [53]	0.292	9.12	1.15	3.04	0.696	0.090
IPAdapter [76]	0.260	4.18	0.63	3.44	**0.815**	**0.057**
InstantStyle [67]	0.277	4.21	0.64	3.41	0.753	0.066
Artist [39]	0.272	10.98	1.18	2.83	0.744	0.077
DEADiff [54]	0.295	10.81	1.14	2.66	0.721	0.089
StyleDrop [62]	0.292	12.16	1.28	2.51	0.717	0.090
DreamStyler [3]	0.301	11.90	1.09	2.53	0.691	0.095
CSGO [74]	0.298	6.43	0.83	3.09	0.716	0.089
SADis (Ours)	**0.302**	**3.19**	**0.53**	**3.51**	0.754	0.066

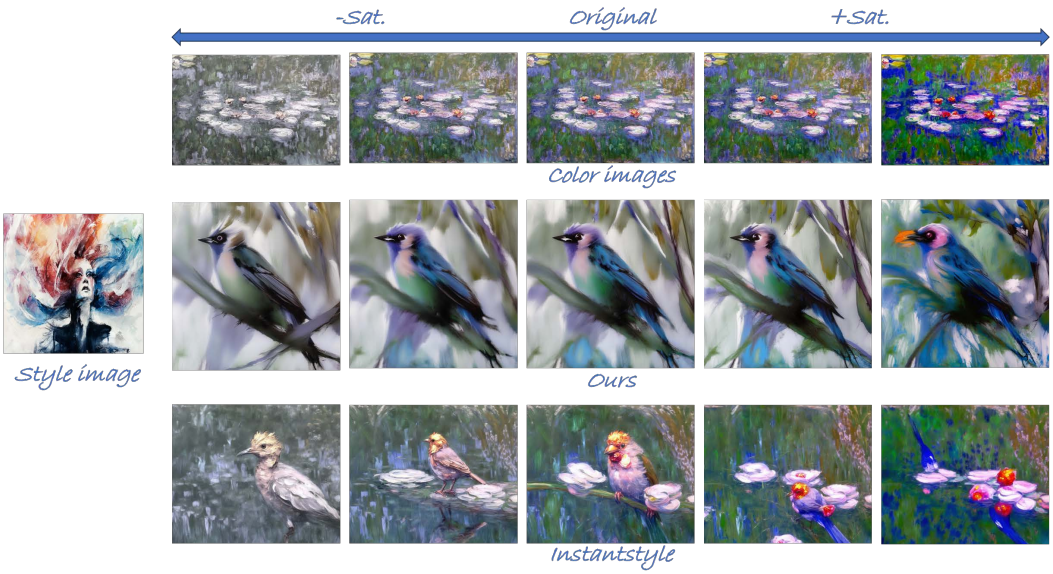

Figure 21: Compared to InstantStyle, *SADis* preserves the content more effectively under saturation variations.

F Ablation studies on *RegWCT*

***RegWCT* scales.** Given a latent $z_t^{'}$ at timestep t, the color rectified latent z_t after *RegWCT* is presented as:

$$z_t = (1-\omega)\, z_t^{'} + \omega \cdot \mathbf{RegWCT}(z_t^{'}). \tag{7}$$

Here, we ablate the balancing factor ω, as shown in Tab. 6. With the increasing scale ω of *RegWCT*, the color score of *SADis* improves significantly, while texture preservation remains largely unaffected. Based on the ablation results in Tab. 6, we set ω to 0.5 to achieve a balanced performance between color and texture alignment.

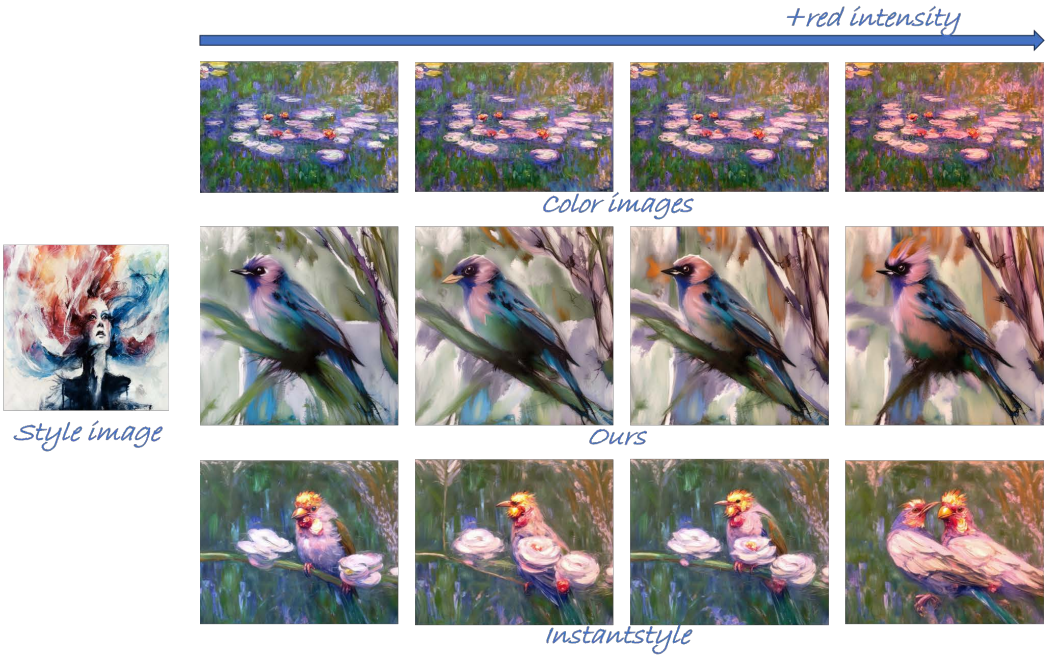

Figure 22: Compared to InstantStyle, *SADis* preserves the content more effectively under diverse illuminant variations.

Table 5: Quantitative results for different sampling rounds.

Round	CLIP	Color			Texture	
		MS-SWD↓	C-Hist↓	GPT4o↑	CLIP-I↑	KID↓
1	0.301	6.08	0.95	3.14	0.751	0.064
2	0.301	6.05	0.95	3.15	0.750	0.065
3	0.300	6.23	0.96	3.12	0.752	0.064
4	0.301	6.16	0.95	3.12	0.751	0.066
5	0.304	6.08	0.95	3.14	0.749	0.066

Timestep intervals. The ablation study on different timesteps to apply *RegWCT* is shown in Fig. 23-(a). During the denoising process, applying *RegWCT* to the latent enhances the color alignment performance but sacrifices texture preservation. As observed in Fig. 23-(a), the early stages of denoising contribute more to the color than the latter stages, which is also supported by previous works [17, 81, 66]. Therefore, we only apply *RegWCT* during the early stages of the denoising, specifically within $[0.8T, 0.6T]$ to achieve a balance between the color and texture generation.

Scale λ of noise injection. The ablation study on the noise scale λ is illustrated in Fig. 23-(b). Without noise injection ($\lambda = 0$), applying WCT improves color alignment but also causes great texture degradation for *DisIG*. This issue is alleviated by injecting the specific degree of latent noise z_T, which is demonstrated in Fig. 23-(b). To achieve a balanced color and texture alignment, we set λ to 0.01 in the experiments based on the results shown in Fig. 23-(b).

G Ablation studies on *CTE*.

Scale γ of SVD. The ablation results of the scaling factor γ and β is presented in Tab. 7 and Tab. 8, respectively. As the scaling factor γ increases, the color scores (such as MS-SWD and C-HIST) improve, with the sacrifice of the slightly texture performance degradation (revealed by CLIP-I). When β increases beyond 1, the texture consistency metric (CLIP-I) improves, indicating enhanced texture alignment. However, this comes at the cost of decreased color consistency, as reflected by the MS-SWD and C-Hist scores. Therefore, to achieve a better balanced performance between texture and color consistency, we set the default values to $\beta = 1$ and $\gamma = 0.003$ in our main experiments.

Table 6: Ablation studies of *RegWCT* scales. considering the trade-off of texture and color alignment, the *RegWCT* scale is set as 0.5 by default.

RegWCT Scale		0	0.3	0.5	0.7	1.0
Color	MS-SWD (↓)	8.05	5.65	5.57	5.23	5.18
	C-Hist (↓)	1.06	0.98	0.96	0.89	0.889
Texture	CLIP-I (↑)	0.747	0.744	0.743	0.741	0.740

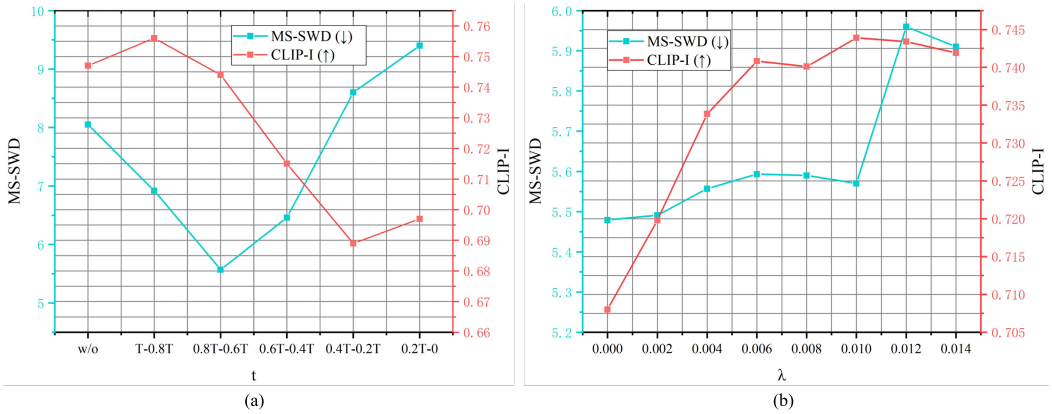

Figure 23: Ablation studies on applying *RegWCT* to different timestep intervals (left). Ablation studies on the scale λ of noise injection (right) during applying *RegWCT*. By default, λ is set to 0.01.

Figure 24: Results of different scales of color embedding $\mathbf{Emb}_{clr}$. Here, the scale of $\mathbf{Emb}_{tx}$ is fixed as 1.

Table 7: Ablation studies of scaling factor γ. γ is set to 0.003 according to the ablation results.

γ ($\beta = 1$)		0	0.001	0.003	0.005	0.007	0.009	0.011
Color	MS-SWD (↓)	5.70	5.71	5.57	5.36	5.22	5.21	5.20
	C-Hist (↓)	1.01	1.017	0.962	0.939	0.897	0.887	0.861
Texture	CLIP-I (↑)	0.759	0.755	0.743	0.736	0.730	0.723	0.713

Table 8: Ablation studies of scaling factor β. β is set to 1 according to the ablation results.

β ($\gamma = 0.003$)		0.5	0.7	0.9	1	1.1	1.3	1.5
Color	MS-SWD (↓)	5.20	5.28	5.45	5.57	5.70	5.92	6.17
	C-Hist (↓)	0.890	0.923	0.949	0.962	0.973	0.992	1.011
Texture	CLIP-I (↑)	0.711	0.729	0.740	0.743	0.747	0.749	0.752

Table 9: Ablation results of controlling different cross-attention layers of *SADis*.

Different CA layers	CLIP	Color			Texture	
		MS-SWD↓	C-Hist↓	GPT4o↑	CLIP-I↑	KID↓
b0a1(ours)	0.301	6.08	0.95	3.14	0.751	0.064
b0a0+b0a1	0.292	6.54	0.981	3.13	0.785	0.057
b0a0+b0a1+b0a2	0.292	6.53	0.982	3.12	0.793	0.056
b0a1+b1a0	0.291	6.55	0.981	3.11	0.791	0.057
b0a1+b1a1	0.300	6.54	0.983	3.12	0.751	0.065
b0a1+b1a2	0.297	7.19	1.005	3.10	0.753	0.066
b0+b1	0.291	7.21	0.998	3.07	0.794	0.055
b0+b1+b2	0.278	7.24	1.003	3.08	0.807	0.056

Color-texture scales. As shown in Fig. 24, we fix the scale of the texture embedding $\mathbf{Emb}_{tx}$ to 1 while varying the scale of the color embedding $\mathbf{Emb}_{clr}$ from 0 to 1.8. With a greater weight for the color embedding, the color score improves. However, it is worth noting that when the scale of the color embedding is set too high, it may negatively impact texture preservation and introduce artifacts that are not derived from either the color or texture reference images.

H Ablation studies on controlling different cross-attention layers.

The purpose of employing multiple Cross-Attention (CA) layers in IP-Adapter is to ensure consistency in the identity of the input. Prior researches [67, 68] indicate that different CA layers in IP-Adapter control different attributes. For example, some are more associated with content, while others relate more to style. Our decision to inject at the first decoder layer was initially inspired by InstantStyle [67]. Experiments with multiple CA layers are provided in Tab. 9. Here, 'bidx0aidx1' denotes the 'idx1'-th CA layer in the 'idx0'-th decoder block of the denoising UNet. As shown in the table, injecting into additional CA layers beyond the style-relevant one (i.e., the first decoder layer) improves texture metrics but degrades color representation. Moreover, we observed that controlling additional CA layers leads to generated images containing more semantic content from the texture reference image (similar to IP-Adapter's results shown in Fig. 1), which is inconsistent with the goal of texture transfer. This observation is consistent with the findings reported in InstantStyle.