DEADiff: An Efficient Stylization Diffusion Model with Disentangled Representations

Tianhao Qi[1,2] Shancheng Fang[1] Yanze Wu[2] Hongtao Xie[1*] Jiawei Liu[2] Lang Chen[2]
Qian He[2] Yongdong Zhang[1]
[1]University of Science and Technology of China [2]ByteDance Inc.
qth@mail.ustc.edu.cn {fangsc, htxie, zyd73}@ustc.edu.cn
{wuyanze.cs, liujiawei.cc22, chenlang.cl, heqian}@bytedance.com

Figure 1. Given a style reference image, **DEADiff** is capable of synthesizing new images that resemble the style and are faithful to text prompts simultaneously. However, previous encoder-based methods (*i.e.*, T2I-Adapter [17]) significantly impair the text controllability of the diffusion-based text-to-image models.

Abstract

The diffusion-based text-to-image model harbors immense potential in transferring reference style. However, current encoder-based approaches significantly impair the text controllability of text-to-image models while transferring styles. In this paper, we introduce DEADiff to address this issue using the following two strategies: 1) a mechanism to decouple the style and semantics of reference images. The decoupled feature representations are first extracted by Q-Formers which are instructed by different text descriptions. Then they are injected into mutually exclusive subsets of cross-attention layers for better disentanglement. 2) A non-reconstructive learning method. The Q-Formers are trained using paired images rather than the identical target, in which the reference image and the ground-truth image are with the same style or semantics. We show that DEADiff attains the best visual stylization results and optimal balance between the text controllability inherent in the text-to-image model and style similarity to the reference image, as demonstrated both quantitatively and qualitatively. Our project page is https://tianhao-qi.github.io/DEADiff/.

*Corresponding author.

1. Introduction

Recently, Diffusion models [21, 22, 25] in text-to-image generation have sparked widespread research due to their astounding performance. As Diffusion models are notoriously known for lacking enhanced controllability, how to stably and reliably guide them to adhere to a predetermined style defined by a reference image becomes intractable.

Taking into account both effectiveness and efficiency, a prevalent method for style transferring is the approach centered around an additional encoder [10, 14, 17, 32, 34, 38]. The encoder-based methods typically train an encoder to encode a reference image to informative features, which are then injected into the Diffusion model as its guided condition. Note that the encoder-based methods are quite efficient due to a single-pass computation, compared with the optimization-based methods that require multiple-iteration learning [5, 9, 13, 24, 27, 37]. Through such an encoder, highly abstract features can be extracted to effectively describe the style of the reference image. These rich style features enable the Diffusion model to accurately understand the style of the reference image it needs to synthesize, as shown on the left side of Fig. 1 where a typical

method (T2I-Adapter [17]) can generate naturally faithful reference styles. However, this approach also introduces a particularly vexing issue: while it allows the model to follow the style of the reference image, it significantly diminishes the model's performance in understanding the semantic context of text conditions.

The loss of text controllability primarily stems from two aspects. On the one hand, the encoder extracts information that couples style with semantics, rather than purely style features. Specifically, previous methods lack an effective mechanism in their encoders to distinguish between image style and image semantics. Therefore the extracted image features inevitably encompass both stylistic and semantic information. This image semantics conflicts with the semantics in the text conditions, leading to a weakened control over text-based conditions. On the other hand, previous methods treat the learning process of the encoder as a reconstruction task, where the ground-truth of the reference image is the image itself. Compared to training a text-to-image model to follow text descriptions, learning from the reconstruction of reference images is typically easier. Consequently, under the reconstruction task, the model tends to focus on the reference image, while neglecting the original text condition in the text-to-image model.

Concerning the above problems, we thus propose *DEADiff* to efficiently transfer reference style to synthetic images without the loss of controllability of text condition. The *DEADiff* consists of two components. Firstly, we decouple the style from the semantics in the reference image from the aspects of feature extraction and feature injection. For feature extraction, a dual decoupling representation extraction mechanism (DDRE) is proposed that utilizes Q-Former [15] to obtain style and semantic representations from the reference image. The Q-Former is instructed by "style" and "content" conditions to selectively extract features that align with the given instructions. For feature injection, we introduce a disentangled conditioning mechanism to inject decoupled representations into mutually exclusive subsets of cross-attention layers for better disentanglement, which is inspired by that different cross-attention layers in the Diffusion U-Net express distinct responses to style and semantics, as demonstrated in [31]. Secondly, we propose a non-reconstruction training paradigm that learns from paired synthetic images. Specifically, the Q-Former instructed by the "style" condition is trained using paired images with the same style as the reference image and the ground-truth image, respectively. Meanwhile, the Q-Former instructed by the "content" condition is trained by images with the same semantics but different styles.

With the style and semantics decoupling mechanism and the non-reconstruction training objective, our *DEADiff* can successfully imitate the style of the reference image, and be faithful to various text prompts, as illustrated in Fig. 1 (b). Compared with the optimization-based methods, our method is more efficient while simultaneously maintaining exceptional style transfer capabilities. In contrast to traditional encoder-based methods, our approach can effectively preserve text control ability. Besides, *DEADiff* eliminates the need for manually adjusting trivial parameters to obtain satisfactory styles, something like feature fusion weight that is typically required by previous methods (*e.g.*, T2I-Adapter).

In summary, our contributions are threefold:

- We propose a dual decoupling representation extraction mechanism to separately obtain style and semantic representations of the reference image, alleviating the problem of semantics conflict between text and reference images from the perspective of learning tasks.
- We introduce a disentangled conditioning mechanism that allows different parts of the cross-attention layers to be responsible for the injection of image style/semantic representation separately, reducing the semantics conflict further from the perspective of model structure.
- We build two paired datasets to aid the DDRE mechanism using the non-reconstruction training paradigm.

2. Related Work

2.1. Diffusion-based Text-to-Image Generation

In recent years, diffusion models have achieved great success in image generation. Diffusion Probabilistic Models (DPMs) [26] are proposed to learn to restore the target data distributions destroyed by the forward diffusion process. DPMs have attracted increasing attention in the community of image synthesis since the initial diffusion-based image generation works [4, 8, 28] prove their powerful generation capacity. Latest diffusion models [21, 22, 25] further achieve state-of-the-art performance on text-to-image generation, which benefits from large-scale pre-training. These methods use U-Net [23] as the diffusion model, in which cross-attention layers are utilized for injecting the text features extracted from the pre-trained encoders [19, 20]. Especially, Latent Diffusion Models (LDMs) [22], which are also known as Stable Diffusion (SD) models, transfer the diffusion process to a low-resolution latent space through a pre-trained auto-encoder and achieve efficient high-resolution text-to-image generation. Considering the great success of diffusion-based text-to-image (T2I) generation models, abundant of recent diffusion methods [14, 34, 35] focus on using more conditions from a reference image. One typical representative is the style, which is the main concern of this paper.

2.2. Stylized Image Generation with T2I Models

Stylized image generation has widely studied based on pre-trained deep convolutional or transformer-based neural net-

works [1–3, 6, 12, 18, 33], which have made substantial advancements, leading to numerous practical applications.

Witnessed by the power of large-scale Text-to-image models, how to utilize these models to fulfill stylized image generation with better quality and more flexibility is an exciting topic to explore. Textual inversion-based methods [5, 37] project the style image into a learnable embedding of the text token space. Unfortunately, the problem of information loss, stemming from the mapping from visual to text modalities, presents a significant challenge to the learned embedding in accurately rendering the style of the reference image with user-defined prompts. In contrast, DreamBooth [24] and Custom Diffusion [13] can synthesize images that better capture the style of the reference image by optimizing all or partial parameters of the diffusion model. Nevertheless, the cost is the decreased fidelity to text prompts resulting from the severe overfitting. Currently, parameter-efficient fine-tuning provides a more effective approach for stylized image generation without impacting the diffusion model's fidelity to text prompts, such as InST [37], LoRA [9] and StyleDrop [27]. However, while these optimization-based methods can customize styles, they all require minutes to hours to fine-tune the model for each input reference image. The additional computational and storage overhead impedes the practicality of these methods in real-world production.

Thus, some optimization-free methods [10, 17, 32] are proposed to extract style features from the reference image through designed image encoders. Among them, T2I-Adapter-Style [17] and IP-Adapter [34] use Transformer [30] as the image encoder with CLIP [19] image embeddings as input, and utilize the extracted image features through U-Net cross-attention layers. BLIP-Diffusion [14] builds a Q-Former [15] to transform the image embeddings to text embedding space and input them to the text encoder of the diffusion model. Those methods use whole image reconstruction [17, 34] or object reconstruction [14] as the training objective, resulting in both the content and style information being extracted from the reference image. To make the image encoders focus on extracting style features, StyleAdapter [32] and ControlNet-shuffle [35] shuffle the patch or pixel of the reference image and could generate various content with the target style.

3. Method

3.1. Preliminary

SD is a type of latent diffusion model [22], which performs a sequence of gradual denoising operations within the latent space and remaps the denoised latent code into the pixel space, thereby generating the final output image. During the training process, SD initially casts an input image x into a latent code z via a Variational Auto-Encoder [11]. In subsequent stages, the noised latent z_t at timestep t serves as the input for the denoising U-Net ϵ_θ, which undertakes interaction with text prompts c via cross-attention. The supervision for this process is ensured by the following objective:

$$L = \mathbb{E}_{z,c,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, c\right)\|_2^2\right], \qquad (1)$$

where ϵ represents a random noise sampled from the standard Gaussian distribution.

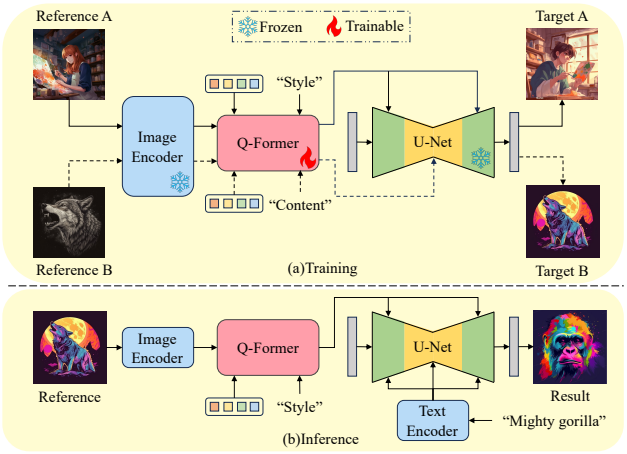

Figure 2. The training and inference paradigm of *DEADiff*. We use proprietary paired datasets for training Q-Former to extract disentangled representations under conditions "style" and "content", which are injected into mutually exclusive cross-attention layers.

3.2. Dual Decoupling Representation Extraction

Taking inspiration from BLIP-Diffusion [14], which learns the subject representations through synthetic image pairs with different background to avoid trivial solution, we integrate two auxiliary tasks that utilize Q-Formers as representation filters nesting within a non-reconstructive paradigm. This enables us to implicitly discern disentangled representations of both style and content within an image.

On the one hand, we sample a pair of distinct images, both maintaining the same style but serving as the reference and target respectively for the Stable Diffusion (SD) generation process, as depicted in pair A of Fig. 2(a). The reference image is fed into the CLIP image encoder, whose output interacts with the learnable query tokens of the Q-Former [15] and its input text through cross-attention. For this process, we settle on the word "style" as the input text in anticipation of generating text-aligned image features as output. This output, which encapsulates the style information, is then coupled with the caption detailing the content of the target image and provided for conditioning to the denoising U-Net. The impetus for this prompt composition strategy aims to better disentangle the style from the content caption allowing the Q-Former to focus more on the extraction of style-centric representations. This learning task is

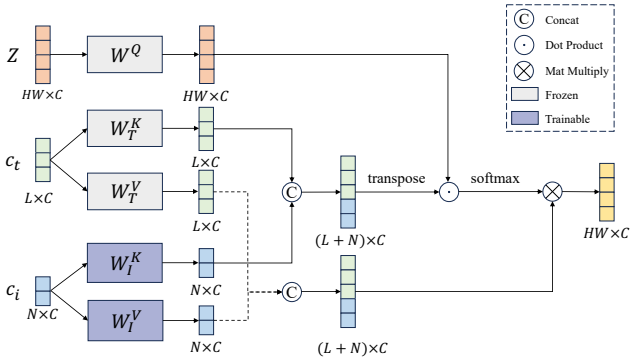

Figure 3. The illustration of our proposed joint text-image cross-attention layer.

defined as the style representation extraction, abbreviated as STRE.

On the other hand, we incorporate a corresponding and symmetric content representation extraction task, referred to as SERE. As shown in pair B of Fig. 2(a), we select two images that share the same subject matter but exhibit distinct styles, which are assigned as the reference and target images. Importantly, we replace the input text of the Q-Former with the word "content" to extract associated content-specific representations. To acquire unadulterated content representations, we supply the output of the query token by the Q-Former and the text style words of the target image, concurrently, as the conditioning for the denoising U-Net. In this approach, the Q-Former will sieve out content unrelated information nested within the CLIP image embeddings while generating the target image.

Simultaneously, we incorporate a reconstruction task into the entire pipeline. The conditioning prompt consists of the query tokens processed by the "style" Q-Former and "content" Q-Former for this learning task. In this way, we can ensure that Q-Formers do not neglect essential image information, considering the complementary relationship between content and style.

3.3. Disentangled Conditioning Mechanism

Motivated by the observation in [31] that different cross-attention layers in the denoising U-Net dominate different attributes of the synthesized image, we introduce an innovative Disentangled Conditioning Mechanism (DCM). In essence, DCM adopts a strategy that conditions the coarse layers with lower spatial resolution on semantics, while the fine layers with higher spatial resolution are conditioned on the style. As illustrated in Fig. 2(a), we only inject the output queries of the Q-Former with "style" conditions to fine layers, which respond to local area features rather than global semantics. This structural adaptation propels the Q-Former to extract more style-oriented features, such as strokes, textures, and colors of the image when inputted with "style" conditions, while diminishing its focus on global semantics. This strategy hence enables a more effective decoupling of style and semantic features. Simultaneously, to make the denoising U-Net support image features as conditions, we devise a joint text-image cross-attention layer, as demonstrated in Fig. 3. In a manner akin to IP-Adapter [34], we include two trainable linear projection layers W_I^K, W_I^V to process image features c_i, in conjunction with frozen ones W_T^K, W_T^V for text features c_t. However, instead of executing cross-attention for image and text features independently, we concatenate the key and value matrices from text and image features respectively, subsequently initiating a single cross-attention operation with U-Net query features Z. Formally, the formulation of this combined text-image cross-attention process can be expressed as follows:

$$Q = ZW^Q, \tag{2}$$

$$K = Concat(c_t W_T^K, c_i W_I^K), \tag{3}$$

$$V = Concat(c_t W_T^V, c_i W_I^V), \tag{4}$$

$$Z^{new} = Softmax(\frac{QK^T}{\sqrt{d}})V. \tag{5}$$

3.4. Paired Datasets Construction

Preparing a pair of images with the same style or subject as stated in Sec. 3.2 is a non-trivial work. Fortunately, existing state-of-the-art text-to-image models have demonstrated a strong fidelity to given text prompts. Therefore, we manually create a list of text prompts by combining subject words and style words, and utilize a pre-trained model to construct two paired image datasets - one with samples of the same style and the other with samples of the same subject. Formally, the construction of the paired datasets involves the following three steps:

Step 1: Text prompt combination. We have listed around 12,000 subject words that span across four major categories: characters, animals, objects, and scenes. Additionally, we have noted more than 650 style words that include attributes such as artistic styles, artists, brushstrokes, shadows, shots, resolutions, and visual angles. Then, every subject word is assigned approximately 14 style words on average from all style words, and the combination forms the final text prompts used for the text-to-image model.

Step 2: Image generation and collection. After combining text prompts with subject words and style words, we have obtained over 160 thousand prompts. Subsequently, all the text prompts are sent to Midjourney, a leading text-to-image generation product, to synthesize corresponding images. As a characteristic of Midjourney, the direct output of a given prompt embraces 4 images with resolution 512×512. We upsample each image to resolution 1024×1024 and store it with the given prompt. Due to redundancy in data collection, we ultimately collected a total of 1.06 million image-text pairs.

Step 3: Paired images selection. We observe that even with the same style words, there are significant differences in images generated with different subject words. In light of this, for the style representation learning task, we use two distinct images synthesized with the same prompt, which serve as the reference and target respectively, as illustrated in Figure Fig. 2(a). To achieve this goal, we store images with the same prompt as a single item and randomly select two images during each iteration. In terms of the content representation learning task depicted in Fig. 2(b), we pair images with the same subject word but different style words as a single item. Ultimately, we have obtained one dataset with over 160000 items for the former task and another one with 1.06 million items for the latter task.

3.5. Training and Inference.

We employ the loss function depicted in Eq. (1) to supervise the aforementioned three learning tasks. During the training process, only the Q-Former and the newly added linear projection layers are optimized. The inference process is illustrated as shown in Fig. 2(b).

4. Experiment

4.1. Experiment Settings

Implementation Details. We adopt Stable Diffusion v1.5 as our base text-to-image model, which comprises a total of 16 cross-attention layers. We number them from 0 to 15 in the order from input to output and define layers 4-8 as coarse layers that are used for injecting image content representation. Accordingly, the other layers are defined as fine layers used for injecting image style representation. We utilize ViT-L/14 from CLIP [19] as the image encoder and keep the number of learnable query tokens of the Q-Former consistent with BLIP-Diffusion, *i.e.*, 16. We adopt two Q-Formers to separately extract semantic and style representations, to encourage them to focus on their own tasks. For the sake of fast convergence, we initialize the Q-Former with the pre-trained model provided by BLIP-Diffusion [14] in HuggingFace[1]. In terms of the additional projection layers W_I^K, W_I^V, we initialize them with the parameters of W_T^K, W_T^V. During training, we set the sampling ratio of the three learning tasks as stated in Sec. 3.2 to 1:1:1, to train the style Q-Former and content Q-Former equally. We fix the parameters of the image encoder, text encoder, and original U-Net[23], and only update the parameters of the Q-Former, 16 learnable queries, and the additional projection layers W_I^K, W_I^V. The models are trained with a total batch size of 512 on 16 A100-80G GPUs. We employ AdamW [16] as the optimizer with a learning rate of $1e-4$ and train for 100000 iterations. As for inference, we adopt the DDIM [28] sampler with 50 steps. The guidance scale for classifier-free guidance [7] is 8.

Datasets. We use self-constructed datasets as introduced in Sec. 3.4 to train our model. The initial dataset with 1.06 million image-text pairs is prepared for the reconstruction task. The style representation learning task is trained using 160000 pairs of images with the same style, while the semantic representation learning task is trained using 1.06 million pairs of images with the same semantics. Please refer to the supplementary material for more detailed information about self-constructed datasets. To evaluate the effectiveness of DEADiff, we construct an evaluation set comprising 32 style images collected from the WikiArt dataset [29] and the Civitai platform. We exclude text prompts with redundant subjects released in StyleAdapter [32], slimming down the original 52 to a final 35. We follow the practice of StyleAdapter, employing Stable Diffusion v1.5 to generate content images corresponding to these 35 text prompts, facilitating comparison with style transfer methods, such as CAST [36] and StyleTR^2 [3].

Evaluation Metrics. In the absence of a precise and suitable metric for assessing style similarity (SS), we propose a more reasonable approach as elaborated in Sec. 7.1. Additionally, we determine the cosine similarity within the CLIP text-image embedding space between the textual prompts and their corresponding synthesized images, indicative of the text alignment capability (TA). We also report the results for the image quality (IQ) of each method. Finally, to eliminate the interference caused by randomness in the objective metric calculation, we conduct a user study to reflect the subjective preference (SP) for the results.

4.2. Comparison with State-of-the-Arts

In this section, we compare our method with the state-of-the-art methods, including optimization-free approaches such as CAST[36], StyleTr^2[3], T2I-Adapter[17], IP-Adapter[34] and StyleAdapter[32], as well as optimization-based methods like InST[37]. It should be noted that since StyleAdapter is not open-sourced, we directly use the results from its released paper for demonstration.

Qualitative Comparisons. Fig. 4 illustrates the comparison results with the state-of-the-art methods. From this figure, we can discern several noteworthy observations. Firstly, the content image-based style transfer methods, such as CAST [36] and StyleTr^2 [3], which do not leverage diffusion models, bypass the issue of reduced text control. However, they merely execute the straightforward color transfer and refrain from engaging more distinctive features like brush strokes and textures from the reference image, leading to noticeable artifacts in each synthesized outcome. Consequently, when such methods encounter scenarios with intricate style references and sizable complex-

[1] https://huggingface.co/salesforce/blipdiffusion

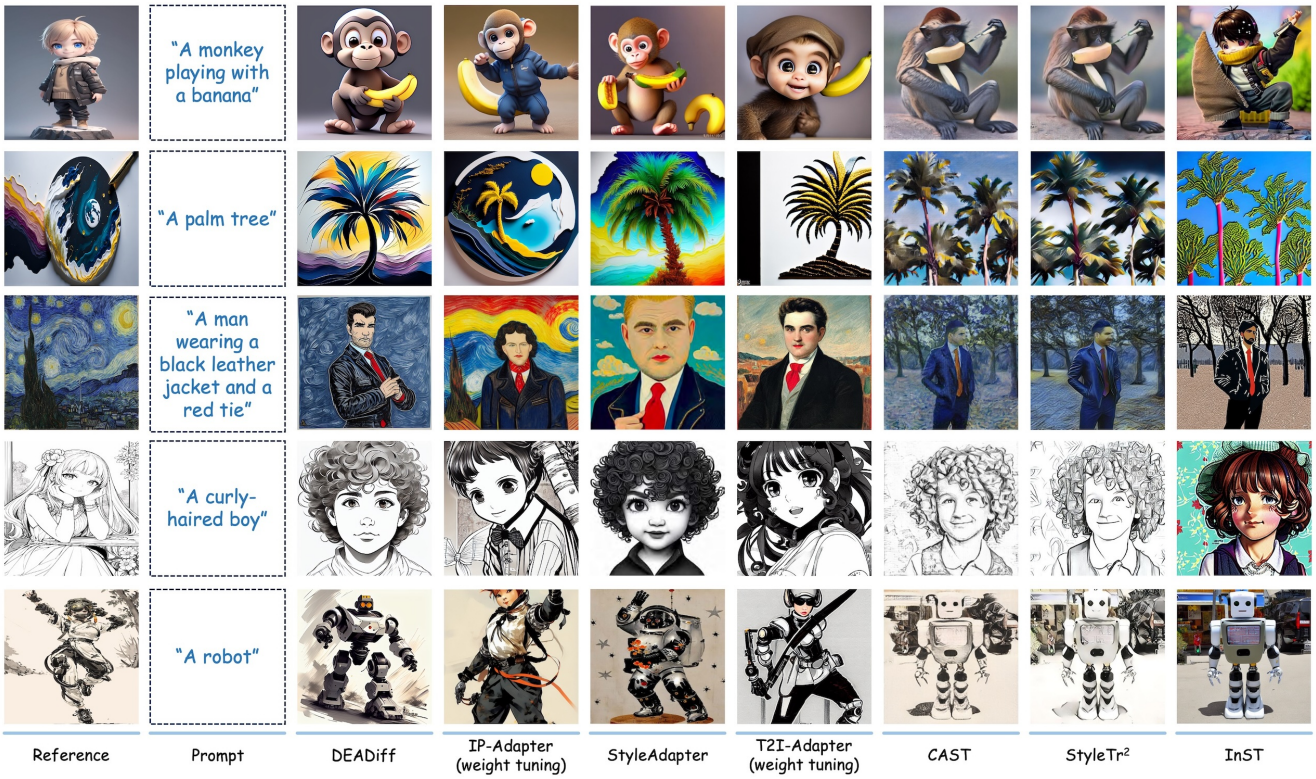

Figure 4. Qualitative comparison with the state-of-the-art methods. Zoom in for better visualization.

ity in content image structures, their style transfer ability notably diminishes. Additionally, for methods trained with the reconstruction objective utilizing diffusion models, whether they are optimization-based (InST [37]) or optimization-free (T2I-Adapter [17]), they generally face semantics interference from the style images in the generated results, as shown in the first and fourth rows of Fig. 4. This aligns with our previous analysis of the semantics conflict issue. Thirdly, while the subsequent improved work, StyleAdapter [32], effectively tackles the problem of semantics conflicts, the style it learns is suboptimal. It loses the detailed strokes and textures of the reference, and there are also noticeable differences in color. Lastly, IP-Adapter [34] with meticulous weight tuning for each reference image can achieve decent results, but its synthesized outputs either introduce some semantics from the reference images or suffer from style degradation. On the contrary, our method not only better adheres to the textual prompts but also significantly preserves the overall style and detailed textures of the reference image, with very minor differences in the color tones.

Quantitative Comparisons. Tab. 1 presents the style similarity, image quality, text alignment and the overall subjective preference of our method compared with the state-of-the-art methods on the evaluation set we constructed. We draw several conclusions from this table. First, aside from T2I-Adapter [17] and IP-Adapter [34] without meticulous weight tuning (whose generated results are often a reorganization of the reference images, as evidenced by their low text alignment scores), we achieve the highest style similar-

Method	SS↑	IQ↑	TA↑	SP↑
InST [37]	0.215	5.148	0.237	6.3
CAST [36]	0.224	4.922	<u>0.282</u>	8.7
StyTr^2 [3]	0.214	5.037	<u>0.282</u>	<u>13.1</u>
T2I-Adapter [17]	0.241	5.500	0.224	2.7
IP-Adapter [34]	**0.274**	<u>5.598</u>	0.155	-
DEADiff	<u>0.229</u>	**5.840**	**0.284**	**69.0**

Table 1. Quantitative comparison with the state-of-the-arts.

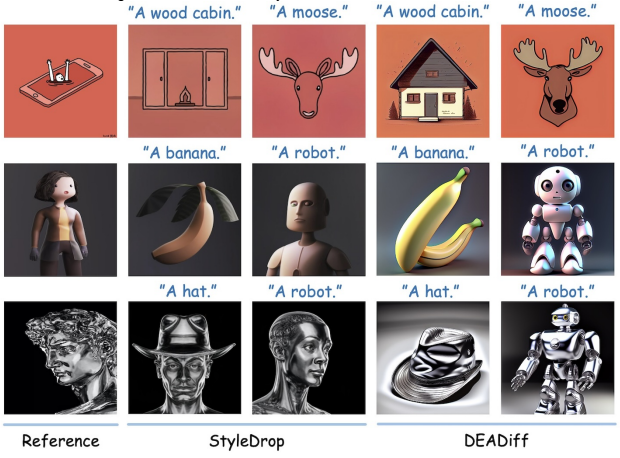

Figure 5. Visual comparison between StyleDrop and *DEADiff*.

Figure 6. Representative visual results under all configurations listed in Tab. 2.

Method	Style Similarity↑	Text Alignment↑
Baseline	**0.274**	0.148
+ DCM	0.259	0.224
+ STRE	0.222	0.286
+ SERE	0.221	0.287
DEADiff	0.224	**0.289**

Table 2. Quantitative results from gradually increasing components with *DEADiff*.

ity, demonstrating that our method indeed effectively captures the overall style of the reference images to some extent. Second, our method achieves comparable text alignment to the two SD-based methods for generating content images, CAST [36] and StyTr^2 [3]. This indicates that our method does not compromise the original text control capabilities of SD while learning the style of the reference images. Third, the substantial advantage reflected in the image quality metric compared to all other methods corroborates the practicality of our approach. Furthermore, as shown in the rightmost column of Tab. 1, users demonstrate a significantly greater preference for our method over all other ones. More detailed results and explanations could be found in supplement materials Sec. 7.1 and Sec. 7.2. In summary, *DEADiff* achieves an optimal balance between text fidelity and image similarity with the most pleasing image quality.

Comparison with StyleDrop [27] Additionally, Fig. 5 presents a visual comparison between our method and StyleDrop. Overall, although *DEADiff* is slightly inferior to optimization-based StyleDrop in terms of color accuracy, it achieves comparable or even better results in terms of artistic style and fidelity to the text. The cabin, hat, and robot generated by *DEADiff* are more appropriate and do not suffer from semantic interference inherently present in the reference image. This demonstrates the critical role of disentangling semantics from the reference image.

4.3. Ablation Study

To comprehend the roles each component plays within *DEADiff*, we conduct a series of ablation studies. Tab. 2 presents the quantitative results under all configurations, whereas Fig. 6 enumerates representative visual outcomes. Note that the baseline refers to injecting image features extracted by Q-Former into all cross-attention layers of the U-Net [23], which is trained with the reconstruction paradigm. Each configuration is assessed on the evaluation set after training 50,000 iterations.

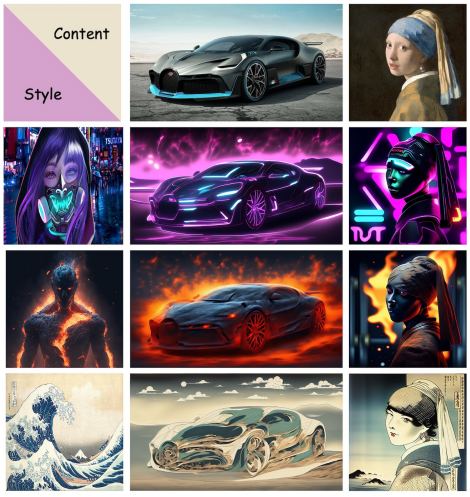

Figure 7. Visual results for content image-based stylization.

Disentangled Conditioning Mechanism. Combining the top two rows of Tab. 2 and the second and third columns of Fig. 6, it is clear that the reconstruction training paradigm inevitably introduces semantics from the reference image, masking the control capabilities of text prompts. Even though DCM does enhance it by capitalizing on U-Net's characteristic of responding differently to conditions at different layers, as evidenced by the visual results and higher text alignment, the semantic component from image features still conflicts with text semantics.

Dual Decoupling Representation Extraction. Referring to the bottom three rows of Tab. 2 and the rightmost three columns of Fig. 6, we observe a notable enhancement in text editability compared to the former DCM and further progressive improvement. Specifically, STRE (the third row in Tab. 2) introduces a non-reconstructive training paradigm, allowing the features extracted by Q-Former to focus more on the style information of the reference image, thereby reducing the semantic components contained within. Hence, the content of the reference image immediately disappears from the generated results, as depicted in the fourth column in Fig. 6. In addition, while the introduction of SERE (the penultimate row in Tab. 2) seems to have limited impact on the results, its combination with STRE (the last row in Tab. 2) to reconstruct the original image ensures that the extracted two representations are decoupled, complementing each other without omissions. As shown by the last column in Fig. 6, the text control capabilities are perfectly manifested while fully replicating the style of the reference image with the overall *DEADiff*.

4.4. Applications

Combination with ControlNet [35]. *DEADiff* supports all types of ControlNets native to SD v1.5. Taking depth Con-

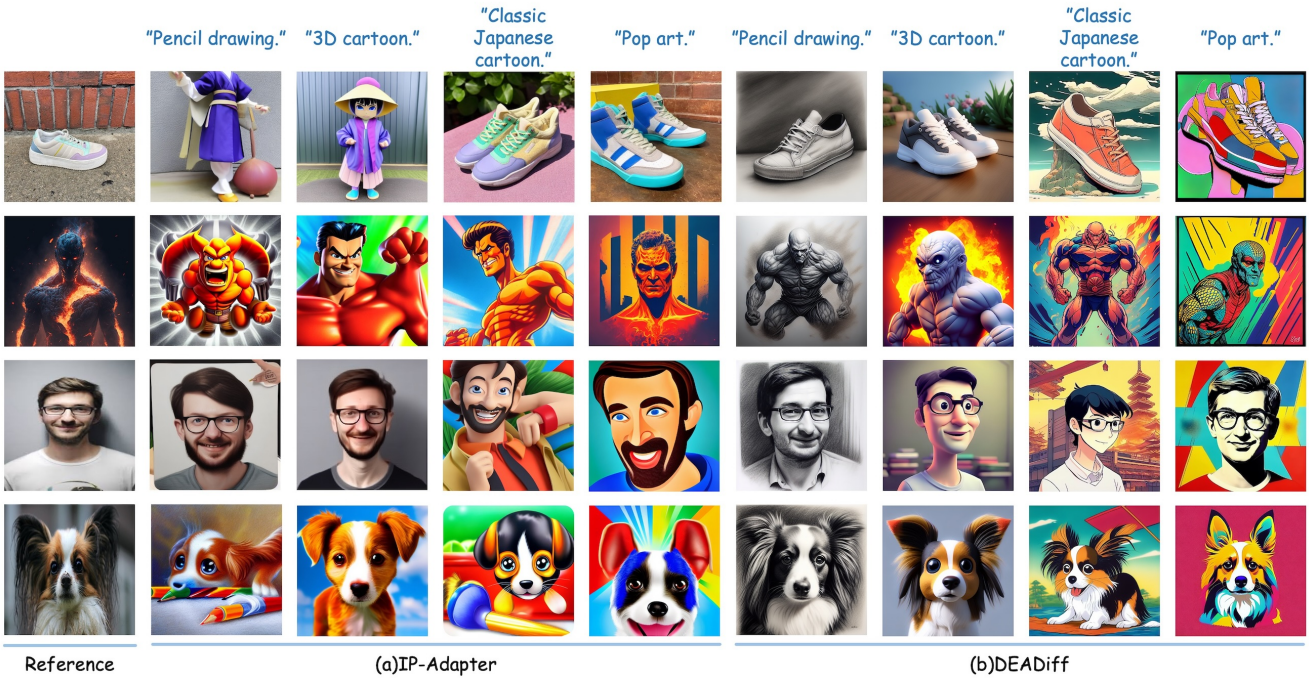

Figure 8. Visual results for the stylization of reference semantics. Note that we reduce the weight of the image condition in IP-Adapter [34] to enhance the efficacy of text prompts in controlling style.

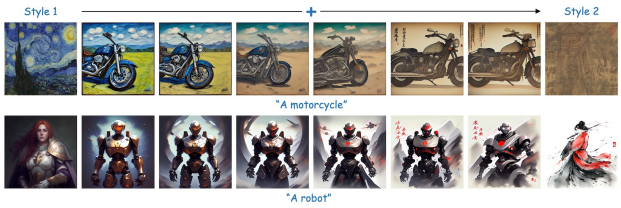

Figure 9. Visual results for style mixing.

Figure 10. Visual results for substituting the denoising U-Net.

trolNet as an example, Fig. 7 demonstrates the impressive effects of stylization while maintaining the layout. *DEADiff* has a wide application scope. In this section, we enumerate a few of its typical applications.

Stylization of reference semantics. Since *DEADiff* can extract the semantic representation of the reference image, it can stylize the semantic objects in the reference image through text prompts. As shown in Fig. 8, the stylization effects are significantly superior to that of IP-Adapter [34].

Style mixing. *DEADiff* is capable of blending styles from multiple reference images. Fig. 9 shows its progressive changing effects under the different control exerted by two reference images.

Switch of the base T2I model. Since *DEADiff* does not optimize the base T2I models, it can directly switch between different base models to generate different stylization results, as shown in Fig. 10.

5. Conclusion

In this paper, we delve into the reasons for the decline in text control capabilities of existing encoder-based stylized diffusion models and subsequently propose the targeted design of *DEADiff*. It includes a dual decoupling representation extraction mechanism and a disentangled conditioning mechanism. Empirical evidence demonstrates that *DEADiff* is capable of attaining an optimal equilibrium between stylization capabilities and text control. Future work could aim to further enhance style similarity and decouple instance-level semantic information.

6. Acknowledgement

This work is supported by the National Nature Science Foundation of China (U23B2028, 62121002, 62102384). We thank our colleagues at ByteDance, Wei Liu and Zhuowei Chen, for their valuable help in this research.

References

[1] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 862–871, 2021. 3

[2] Haibo Chen, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, Dongming Lu, et al. Artistic style transfer with internal-external learning and contrastive learning. *Advances in Neural Information Processing Systems*, 34:26561–26573, 2021.

[3] 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. 3, 5, 6, 7, 1, 2

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[5] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1, 3

[6] 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. 3

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1, 3

[10] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 1, 3

[11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[12] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019. 3

[13] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1, 3

[14] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 1, 2, 3, 5

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 3

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 5

[17] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 1, 2, 3, 5, 6

[18] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5880–5888, 2019. 3

[19] 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. 2, 3, 5

[20] Colin Raffel, Minh-Thang Luong, Peter J Liu, Ron J Weiss, and Douglas Eck. Online and linear-time attention by enforcing monotonic alignments. In *International Conference on Machine Learning*, pages 2837–2846, 2017. 2

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

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2, 5, 7

[24] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1, 3

[25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 2

[26] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265, 2015. 2

[27] 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. *arXiv preprint arXiv:2306.00983*, 2023. 1, 3, 7

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 5

[29] Wei Ren Tan, Chee Seng Chan, Hernan E Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1):394–409, 2018. 5

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[31] 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. 2, 4

[32] 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. *arXiv preprint arXiv:2309.01770*, 2023. 1, 3, 5, 6

[33] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. Styleformer: Real-time arbitrary style transfer via parametric style composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14618–14627, 2021. 3

[34] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 2, 3, 4, 5, 6, 8

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

[36] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–8, 2022. 5, 6, 7, 1, 2

[37] 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. 1, 3, 5, 6, 2

[38] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *arXiv preprint arXiv:2305.16322*, 2023. 1