DreamStyler: Paint by Style Inversion with Text-to-Image Diffusion Models

Namhyuk Ahn[1], **Junsoo Lee**[1], **Chunggi Lee**[1,2], **Kunhee Kim**[3], **Daesik Kim**[1], **Seung-Hun Nam**[1], **Kibeom Hong**[4]

[1] NAVER WEBTOON AI
[2] Harvard University
[3] KAIST
[4] SwatchOn

Abstract

Recent progresses in large-scale text-to-image models have yielded remarkable accomplishments, finding various applications in art domain. However, expressing unique characteristics of an artwork (*e.g.* brushwork, colortone, or composition) with text prompts alone may encounter limitations due to the inherent constraints of verbal description. To this end, we introduce DreamStyler, a novel framework designed for artistic image synthesis, proficient in both text-to-image synthesis and style transfer. DreamStyler optimizes a multi-stage textual embedding with a context-aware text prompt, resulting in prominent image quality. In addition, with content and style guidance, DreamStyler exhibits flexibility to accommodate a range of style references. Experimental results demonstrate its superior performance across multiple scenarios, suggesting its promising potential in artistic product creation. Project page: https://nmhkahn.github.io/dreamstyler/.

Introduction

"*Painting is silent poetry.*" — Simonides, Greek poet

Recent text-to-image models have shown unprecedented proficiency in translating natural language into compelling visual imagery (Saharia et al. 2022; Ramesh et al. 2022; Rombach et al. 2022). These have emerged in the realm of art, providing inspiration and even assisting in crafting tangible art pieces. In the AI-assisted art production workflow, artists typically utilize various descriptive prompts that depict the style and context to generate their desired image. However, the unique styles of a painting, its intricate brushwork, light, colortone, or composition, cannot be easily described in a single word. For instance, dare we simplify the entirety of Vincent Van Gogh's lifelong artworks as just one word, 'Gogh style'? Text descriptions cannot fully evoke his unique style in our imagination — his vibrant color, dramatic light, and rough yet vigorous brushwork.

Beyond text description, recent studies (Gal et al. 2022; Ruiz et al. 2023) embed specific attributes of input images into latent space. While they effectively encapsulate a novel *object*, we observed that they struggle to personalize *style* of a painting. For instance, model optimization-based methods (Ruiz et al. 2023; Kumari et al. 2023) are highly susceptible to overfitting and often neglect inference prompts, which is not ideal for real-world production (please refer to the Suppl. for more details). Textual inversion-based methods (Gal et al. 2022; Voynov et al. 2023), in contrast, effectively reflect the inference prompt but fail to replicate style, possibly due to the limited capacity of the learned embeddings. This is because capturing style, from global elements (*e.g.* colortone) to local details (*e.g.* detailed texture), is challenging when relying solely on a single embedding token.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

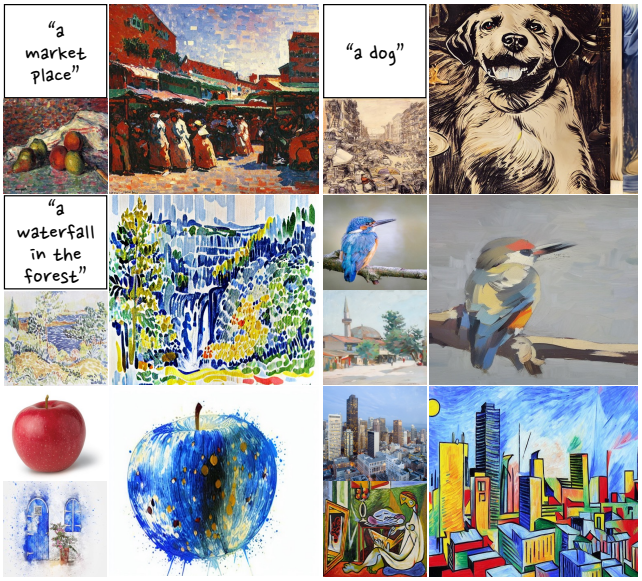

Figure 1: DreamStyler synthesizes outputs based on a given context along with a style reference. Note that each model is trained on a single style image shown in this figure.

In this work, we present **DreamStyler**, a novel single (one-shot) reference-guided artistic image synthesis framework designed for the text-to-image generation and style transfer tasks (Figure 1). We encapsulate the intricate styles of artworks into CLIP text space. DreamStyler is grounded in textual inversion (TI), chosen for the inherent flexibility that stems from its prompt-based configuration. To overcome the limitations of TI, we introduce an extended textual embedding space, $\mathcal{S}$ by expanding textual embedding into the denoising timestep domain (Figure 2). Based on this space, we propose a multi-stage TI, which maps the textual

information into the $\mathcal{S}$ space. It accomplishes by segmenting the entire diffusion process into multiple *stages* (a chunk of timesteps) and allocating each textual embedding vector to the corresponding stage. The exploitation of the timestep domain in textual inversion significantly improves the overall efficacy of artistic image synthesis. This enhancement stems from the increased capacity of the personalized module, as well as the utilization of prior knowledge suggesting that different denoising diffusion steps contribute differently to image synthesis (Balaji et al. 2022; Choi et al. 2022).

We further propose a context-aware prompt augmentation that simply yet proficiently decouples the style and context information from the reference image. With our approach, the personalization module can embed style features solely into its textual embeddings, ensuring a more faithful reflection of the reference's style. To further refine the artistic image synthesis, we introduce a style and context guidance, inspired by classifier-free guidance (Ho and Salimans 2022). Our guidance bisects the guidance term into style and context components, enabling individual control. Such a guidance design allows users to tailor the outputs based on their preferences or intricacy of the reference image's style.

We validate the effectiveness of DreamStyler through a broad range of experiments. DreamStyler not only demonstrates advanced artistic image synthesis but also paves the new way of applying text-to-image diffusion models to the realms of artistic image synthesis and style transfer tasks.

Related Work

Personalized text-to-image synthesis. Since the latent-based text conditional generation has been explored (Rombach et al. 2022), following studies (Saharia et al. 2022; Ramesh et al. 2022; Li et al. 2022) have further contributed to enhancing text-to-image synthesis with CLIP (Radford et al. 2021) guidance. Furthermore, Textual inversion (Gal et al. 2022), DreamBooth (Ruiz et al. 2023) and CustomDiffusion (Kumari et al. 2023) introduced approaches that leverage 3-5 images of the subject to personalize semantic features. Recently, Voynov et al. (2023) proposed $\mathcal{P}+$ space, which consists of multiple textual conditions, derived from per-layer prompts. Although they showed promising results in penalization of diffusion models, there are still limitations to fully capturing precise artistic style representations. In contrast, DreamStyler considers the denoising timestep to accommodate temporal dynamics in the diffusion process, achieving high-quality artistic image generation.

Paint by style. Neural style transfer renders the context of a source with a style image. Since Gatys, Ecker, and Bethge (2016), studies have been devoted to enhancing the transfer networks for more accurate and convincing style transfer. Notably, AdaIN (Huang and Belongie 2017) and AdaAttN (Liu et al. 2021) investigated matching the second-order statistics of content and style images. AesPA-Net (Hong et al. 2023) and StyTr^2 (Deng et al. 2022) adopted recent architectures such as attention or transformer for high-fidelity neural style transfer. Recently, InST (Zhang et al. 2023) utilized the diffusion models by introducing the image encoder to inverse style images into CLIP spaces.

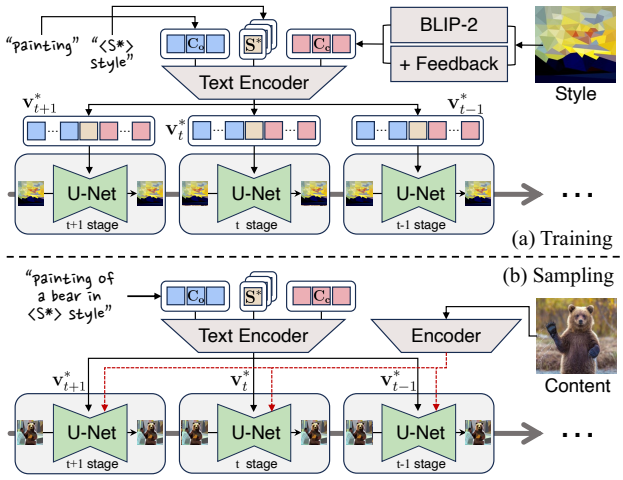

Figure 2: Model overview. (a) DreamStyler constructs training prompt with an opening text C_o, multi-stage style tokens $\mathbf{S}^*$, and a context description C_c, which is captioned with BLIP-2 and human feedback. DreamStyler projects the training prompt into multi-stage textual embeddings $\mathbf{v}^* = \{v_1^*, \ldots, v_T^*\}$, where T is # stages (a chunk of the denoising timestep). As a result, the denoising U-Net provides distinct textual information at each stage. (b) DreamStyler prepares the textual embedding using a provided inference prompt. For style transfer, DreamStyler employs ControlNet to comprehend the context information from a content image.

Method

Preliminary: Stable Diffusion (SD). DreamStyler is built upon SD (Rombach et al. 2022). SD projects an input image x into a latent code, $z = E(x)$ using an encoder E, while decoder D transforms the latent code back into pixel space, *i.e.* $x' = D(z')$. The diffusion model creates a new latent code z' by conditioning on additional inputs such as a text prompt y. The training objective of SD is defined as:

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

At each timestep t, the denoising network ϵ_θ reconstructs the noised latent code z_t, given the timestep t and a conditioning vector $c(y)$. To generate $c(y)$, each token from a prompt is converted into an embedding vector, which is then passed to the CLIP text encoder (Radford et al. 2021).

Preliminary: Textual Inversion (TI). Gal et al. (2022) proposed a method to personalize a pre-trained text-to-image model by incorporating a novel embedding representing the intended concept. To personalize the concept, they initialize a word token S^* and its corresponding vector v^*, situated in the textual conditioning space $\mathcal{P}$, which is the output of the CLIP text encoder. Instead of altering any weights in SD models, they optimize v^* alone using Eq. (1). To create images of personalized concepts, the inclusion of S^* in the prompts (*e.g.* a photo of S^* dog) is the only required step.

Multi-Stage Textual Inversion

In some cases, TI fails to sufficiently represent the concept due to the inherent capacity limitations associated with us-

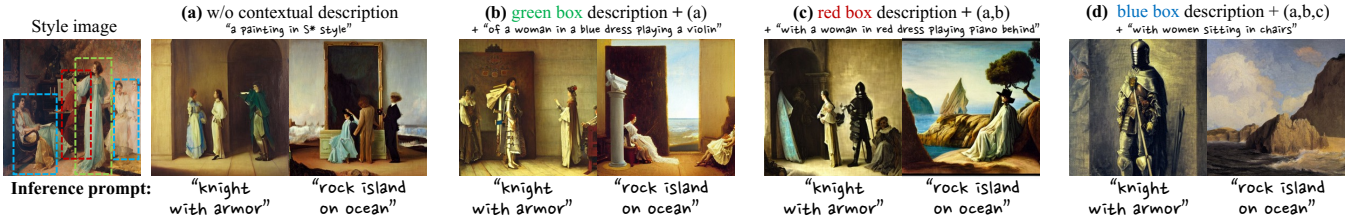

Figure 3: How does training prompt affect? Given a style image, we construct training prompts with contextual descriptions (b∼d). (a) Training without contextual description in the prompt; *i.e.* trains the model with "a painting in S^* style". The model tends to generate the images that contains objects and compositions from the style image (*e.g.* standing and sitting audiences), instead of attributes depicted in the inference prompt. (b, c) Training with partial contextual descriptions (the green and red boxes displayed in the style image, respectively). Such a tendency is significantly reduced, yet the model still synthesizes some objects from the style image (*e.g.* sitting people in the blue box). (d) Training with full contextual descriptions. The model produces outputs that fully reflect the inference prompt without introducing any non-style attributes from the style image.

ing a single embedding token. Moreover, this single embedding strategy is inappropriate for accommodating the changing process of diffusion models. As explored in Balaji et al. (2022); Choi et al. (2022), diffusion models display intriguing temporal dynamics throughout the process, necessitating different capacities at various diffusion steps. In light of this, managing all denoising timesteps with a single embedding potentially has limitations due to the spectrum of local to global expressions embodied in paintings. Thus, articulating paintings is intricately related to the denoising timesteps, which operate in a coarse-to-fine synthesis manner (Balaji et al. 2022). To address these challenges, we introduce a *multi-stage TI* that employs multiple embeddings, each corresponding to specific diffusion stages (Figure 2).

We first propose an extended textual embedding space $\mathcal{S}$. The premise of the $\mathcal{S}$ space is to decompose the entire diffusion process into multiple distinct *stages*. To implement this, we split the denoising timesteps into T chunks and denote each chunk as a stage. Based on the $\mathcal{S}$ space, the multi-stage TI prepares the copies of the initial style token (S^*) as a multi-stage token set $\mathbf{S}^* = \{S_1^*, \dots, S_T^*\}$. In this way, the multi-stage TI projects a style image into T style tokens, contrasting the TI that embeds it into a single token. The token set is then encoded by a CLIP text encoder to form stage-wise embedding vectors, denoted as $\mathbf{v}^* = \{v_1^*, \dots, v_T^*\}$. Lastly, the multi-stage TI optimizes these embeddings following the subsequent equation.

$$\mathbf{v}^* = \arg\min_{\mathbf{v}} \mathbb{E}_{z,\mathbf{v},\epsilon,t}[||\epsilon - \epsilon_\theta(z_t, t, c(v_t))||_2^2]. \quad (2)$$

The application of multi-stage TI significantly enhances the representation capacity beyond that of vanilla TI, which we will illustrate in a series of experiments. Furthermore, this method enables the fusion of multiple tokens, each originating from different styles, at a specific stage t. Consequently, it facilitates the creation of unique and novel styles tailored to the user's individual preferences.

Context-Aware Text Prompt

While the multi-stage TI enhances representational capacity, it still faces fundamental problems when training with a style reference; the style and context of the image may become entangled during the optimization of the embeddings. This problem mainly arises from attempts to encapsulate all features of the image into S^*, not just the style aspect. As depicted in Figure 3, without contextual information in the training prompt, the model overlooks the context of inference prompt. However, when we inject contextual descriptions into the training prompt, the model better disentangles the style from the context. In our observations, such a phenomenon occurs more frequently as the representational capacity increases, likely due to the model's increased efforts to accommodate all information within its capacity.

Hence, we construct training prompts to include contextual information about the style image. Let $C = [C_o, \mathbf{S}^*]$ be the vanilla prompt used in multi-stage TI training, where C_o is the opening text (*e.g.* "a painting"), and $\mathbf{S}^*$ is multi-stage style token set, described above. In the proposed strategy, we incorporate a contextual descriptor C_c (*e.g.* "of a woman in a blue dress") into the middle of the prompt (Figure 2), *i.e.* $C = [C_o, C_c, \mathbf{S}^*]$. We annotate all the non-style attributes (*e.g.* objects, composition, and background) from the style image to form the contextual descriptor. When we caption non-style attributes, BLIP-2 (Li et al. 2023) is employed to aid in the automatic prompt generation.

Although a context-aware prompt significantly reinforces style-context decoupling, for some style images with complicated contexts (Figure 3), BLIP-2 might not capture all details, which could limit the model's disentanglement capability. In such cases, we further refine caption C_c based on human feedback (e.g., caption by humans). This human-in-the-loop strategy is straightforward yet markedly improves the model's ability to disentangle styles. Since our goal is one-shot model training, the time spent refining the caption is minimal; typically less than a minute. With the context-aware prompt, the text-to-image models can now distinguish style elements from contextual ones and specifically embed these into the (multi-stage) style embeddings $\mathbf{v}^*$. The motivation for augmenting the training prompt is also suggested in StyleDrop (Sohn et al. 2023), a current personalization approach in the text-to-image diffusion model.

Style and Context Guidance

Classifier-free guidance (Ho and Salimans 2022) improves conditional image synthesis. It samples adjusted noise pre-

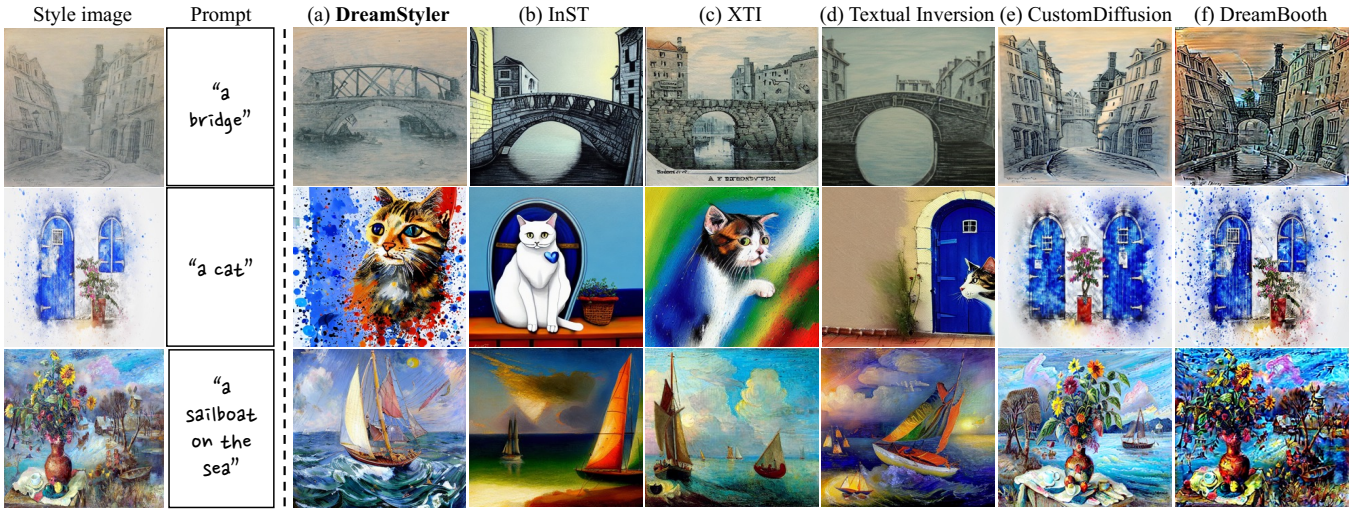

Figure 4: Qualitative comparison on the style-guided text-to-image synthesis task.

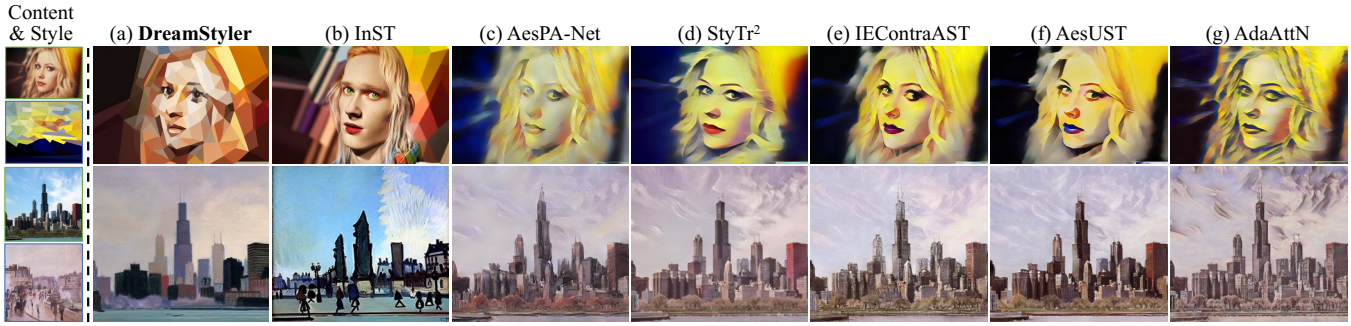

Figure 5: Qualitative comparison on the style transfer task.

diction $\hat{\epsilon}(.)$, by leveraging unconditional output under null token $\emptyset$ as: $\hat{\epsilon}(\mathbf{v}) = \epsilon(\emptyset) + \lambda(\epsilon(\mathbf{v}) - \epsilon(\emptyset))$, where, λ is the guidance scale and we omit $c(.)$, z and t for brevity.

In style-guided image synthesis, this guidance pushes both style and context uniformly with λ. The uniform guidance could face limitations since the spectrum of "style" of artistic paintings is wider than that of natural photos. Given this diversity, a more nuanced control mechanism is required. Furthermore, there exist demands to individually control style and context in the art-making process. To this end, we propose style and context guidance as in below.

$$\begin{aligned}\hat{\epsilon}(\mathbf{v}) = \epsilon(\emptyset) + \lambda_s[\epsilon(\mathbf{v}) - \epsilon(\mathbf{v_c})] + \lambda_c[\epsilon(\mathbf{v_c}) - \epsilon(\emptyset)] \\ + \lambda_c[\epsilon(\mathbf{v}) - \epsilon(\mathbf{v_s})] + \lambda_s[\epsilon(\mathbf{v_s}) - \epsilon(\emptyset)] \quad (3)\end{aligned}$$

where, $\mathbf{v}_s, \mathbf{v}_c$ are the embeddings of prompts C, C_c, respectively. λ_s, λ_c denote style and context guidance scale. We derive Eq. (3) by decomposing $\mathbf{v}$ into $\mathbf{v}_s, \mathbf{v}_c$. We employ two paired terms to balance the influence of each guidance. Please refer to Suppl. for detailed derivation and analysis.

By separating the guidance into style and context, users are afforded the flexibility to control these elements individually. Specifically, an increase in λ_c increases the model's sensitivity towards context (*e.g.* inference prompt or content image), whereas amplifying λ_s leads the model towards a more faithful style reproduction. This flexible design allows users to generate stylistic output tailored to their individual preferences, and it also facilitates the adoption of various styles, each with a range of complexities (Hong et al. 2023).

Style Transfer

DreamStyler transmits styles by inverting a content image into a noisy sample and then denoising it towards the style domain (Meng et al. 2021). With this approach, however, the preservation of content would be suboptimal (Ahn et al. 2023). To improve this, we inject additional conditions from the content image into the model (Zhang and Agrawala 2023) (Figure 2). This straightforward pipeline well preserves with the structure of the content image, while effectively replicating styles. Moreover, by leveraging a powerful prior knowledge from text-to-image models, the style quality of DreamStyler surpasses that of traditional methods.

Experiment

Implementation details. We use $T = 6$ for multi-stage TI and utilize human feedback-based context prompts by default. Please refer to Suppl. for more details.

Datasets. We collected a set of 32 images representing var-

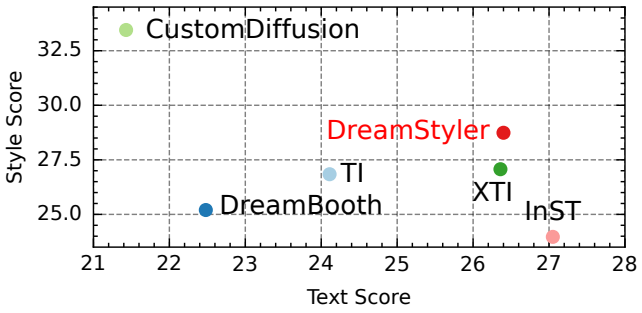

Figure 6: Performance of text and style scores in style-guided text-to-image synthesis. DreamStyler effectively balances these metrics and surpasses the majority of methods.

Method	Text Score	Style Score	User Score
Textual Inversion (Gal et al. 2022)	24.11	26.84	2.1%
DreamBooth (Ruiz et al. 2023)	22.48	25.20	3.9%
CustomDiffusion (Kumari et al. 2023)	21.43	**33.45**	4.8%
XTI (Voynov et al. 2023)	26.36	27.07	4.5%
InST (Zhang et al. 2023)	**27.05**	23.97	1.8%
DreamStyler (Ours)	26.40	28.74	**82.9%**

Table 1: Quantitative comparison on the style-guided text-to-image synthesis task. Bold: best, underline: second best.

ious artistic styles, following the literature on style transfer (Tan et al. 2019). To evaluate text-to-image synthesis, we prepared 40 text prompts, as described in Suppl.

Baselines. In terms of text-to-image synthesis, we compare DreamStyler against diffusion-based personalized methods, ranging from textual inversion to model-optimization approaches. For the style transfer task, we compare our method to state-of-the-art style transfer frameworks. We utilize official codes for all the methods used in the comparison.

Evaluation. Text and image scores, based on CLIP, measure the alignment with a given text prompt and style image, respectively. Style score assesses the style consistency by calculating the similarity of Gram features between the style and generated images. More details are provided in Suppl.

Style-Guided Text-to-Image Synthesis

Table 1 and Figure 6 show quantitative results. DreamStyler delivers a robust performance while managing the trade-off between text and style scores. A tendency is noted that an overemphasis on input text prompts may lead to a compromise in style quality. Despite this, DreamStyler effectively balances these aspects, yielding a performance that goes beyond the trade-off line, indicative of outstanding capability. User score also supports the distinction of DreamStyler.

As shown in Figure 4, previous inversion-based methods (TI, InST, and XTI) effectively preserve the context of text prompts but fall short in adopting the intrinsic artwork of style images. Conversely, the model optimization-based methods (DreamBooth, CustomDiffusion) excel in delivering styles but struggle to adhere to the prompt or introduce

Figure 7: My object in my style. Textual inversion faces challenges in accurately capturing both style and context from the reference images. Although CustomDiffusion successfully recreates the object's appearance, it tends to generate objects in a realistic style, which does not entirely match the target style image. On the other hand, DreamStyler excels at synthesizing the object in the user-specified style.

Method	Text Score	Image Score	User Score
AdaAttN (Liu et al. 2021)	56.67	56.76	8.6%
AesUST (Wang et al. 2022)	58.05	58.09	6.8%
IEContraAST (Chen et al. 2021)	59.38	59.42	8.6%
StyTr^2 (Deng et al. 2022)	56.18	56.28	21.2%
AesPA-Net (Hong et al. 2023)	58.08	58.15	8.6%
InST (Zhang et al. 2023)	65.32	65.37	2.3%
DreamStyler (Ours)	**66.04**	**66.05**	**44.1%**

Table 2: Quantitative comparison on the style transfer task.

objects in style images (3rd row). DreamStyler, in contrast, not only faithfully follows text prompts but also accurately reflects the delicate artistic features of style images.

Style Transfer

As an extended application, DreamStyler also conducts style transfer. As shown in Table 2, we quantitatively compare with previous style transfer studies. Note that since most prior studies have employed Gram loss to boost style quality, we report a CLIP-based image score as an evaluation metric for a more fair comparison. In this benchmark, DreamStyler achieves state-of-the-art performance across text and image scores as well as user preference. Figure 5 also provides evidence of DreamStyler's effectiveness. Our method adeptly captures style features such as polygon shapes or subtle brushwork present in style images. These results highlight the method's capacity to accurately mirror both the thematic intent and the stylistic nuances of the source artwork.

Stylize My Own Object in My Own Style

Beyond style transfer that stylizes *my image*, one might desire to stylize *my object* (Sohn et al. 2023). In such a scenario, a user leverages both their object and style images. As DreamStyler employs an inversion-based approach, this can be readily accomplished by simply training an additional embedding for the object. Subsequently, the user

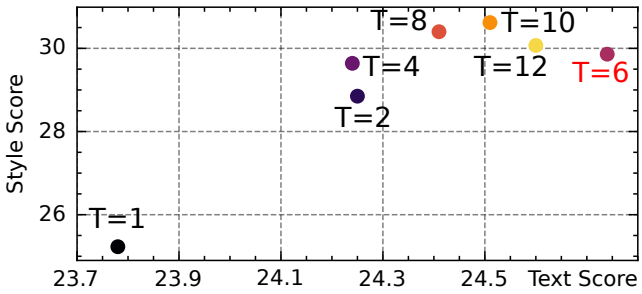

Figure 8: Study on the number of stages (T) in multi-stage TI. We vary T from 1 to 12 and select $T = 6$ as the final model, considering the trade-off between text and style.

Figure 9: Visual comparison of varying T in multi-stage TI. At $T = 1$, the model fails in both style replication and prompt understanding. As T increases, the style quality and text alignment are drastically enhanced.

freely merges style and object tokens in the inference prompt to generate images. As depicted in Figure 7, DreamStyler excels in accurately reflecting both the style and object

Model Analysis

Ablation study. In Table 3, we evaluate each component of our method. The usage of multi-stage TI substantially augments both the text and style score, with a marked increase in style quality, accentuating the pivotal role of this module in creating artistic stylization products. A context-aware prompt yields a modest alteration in the quantitative metrics, yet provides a considerable contribution to the qualitative, which we will discuss in the following section. Style and context (S&C) guidance considerably impacts scores, reinforcing its significance in sustaining the comprehensive quality and coherence of the generated outputs.

Multi-stage TI. In Figure 8, we delve into the influence of the number of stages (T) on performance. A transition from $T = 1$ to 4 results in substantial improvement. Upon reaching $T = 6$, the performance begins to navigate trade-off contours, prompting us to select $T = 6$ for the final model, as we seek to improve the text alignment of the synthesized images. Nevertheless, users have the flexibility to choose a different T value according to their preference. In Figure 9, we provide a visual comparison of the outcomes when T is set to 1, 2, and 6. While $T = 1$ struggles to reflect the artistic features of the style image or comprehend the input prompt,

Method	Text Score	Style Score
Baseline (Gal et al. 2022)	23.78	25.23
+ Multi-Stage TI	24.74	**29.86**
+ Context-Aware Prompt	24.65	29.50
+ S&C Guidance (**Ours**)	**25.38**	29.62

Table 3: Model ablation study. Upon the textual inversion baseline (Gal et al. 2022), we attach the proposed components to measure the effectiveness of our method.

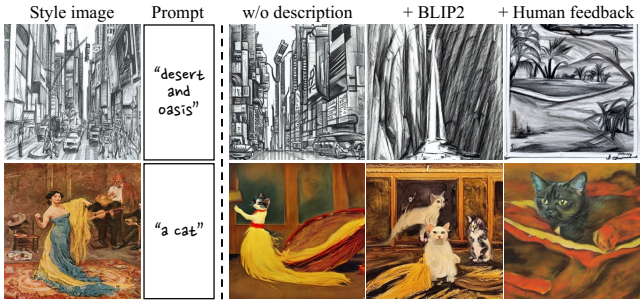

Figure 10: Comparison of three prompt strategies. The model trained without contextual description struggles to disentangle style and context from the style image, generating elements present in the style reference (*e.g.* the same composition in 1st row, a yellow dress in 2nd row). The contextual prompt alleviates this issue to some extent, but the BLIP2-based construction cannot completely eliminate it (*e.g.* the same vanishing point in 1st row). The issue is thoroughly addressed when human feedback is utilized.

$T = 2$ uplifts the quality, yet it also falls short of embracing the style. In contrast, $T = 6$ proves proficient at mimicking the style image, effectively replicating delicate brushwork (1st row) or emulating the pointillism style (2nd row).

Context-aware prompt. Figure 10 presents a visual comparison of three prompt constructions. Training the model without any contextual description (*i.e.* using "A painting in S^* style.") poses a significant challenge, as it struggles to distinguish style from the context within the style image. Subsequently, this often results in the generation of elements that exist in the style reference, such as objects or scene perspective. The introduction of a contextual prompt considerably alleviates this issue, aiding the model in better separating stylistic elements from context. However, the automatic prompt construction does not fully resolve this, as BLIP-based captions often fail to capture all the details of the style image. The most effective solution is leveraging human feedback in the construction of prompts. This approach effectively tackles the issue, resulting in a more robust separation of style and context in the generated outputs.

Guidance. In Figure 11, we explore style and context guidance by adjusting the scale parameters. When we amplified the style guidance strength (λ_s), the model mirrors the style image, illustrating style guidance's capability in managing the image's aesthetics. Yet, overemphasis on style risks compromising the context, leading to outputs that, while stylis-

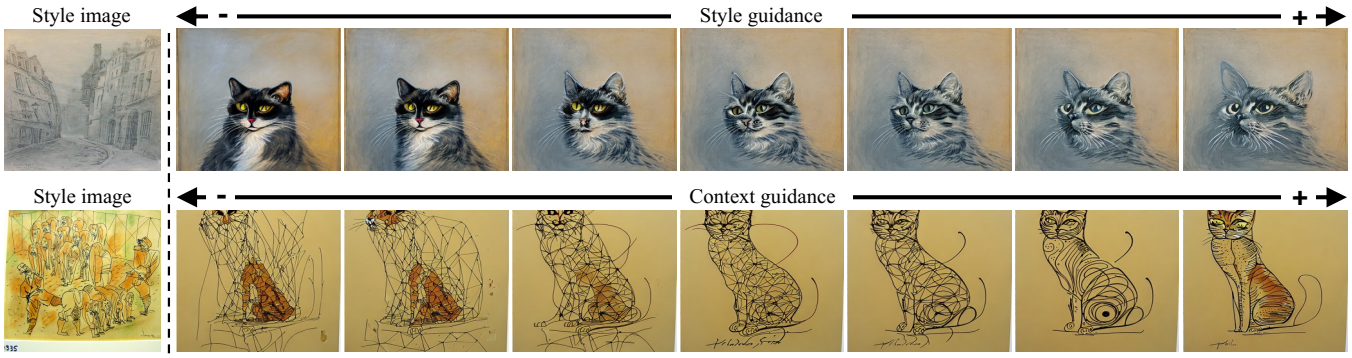

Figure 11: Study on the style and context guidance. Inference prompt: "A cat". By adjusting the scale parameters (λ_s, λ_c), we assess the influence of style and context guidance on the synthesized image. Increasing the style guidance strength causes the model to align more closely with the aesthetics of the style image; however, an excessive emphasis on style could compromise the context. Conversely, increasing the context guidance strength ensures the output corresponds with the inference prompt, but overly strong context guidance could deviate the output from the original style.

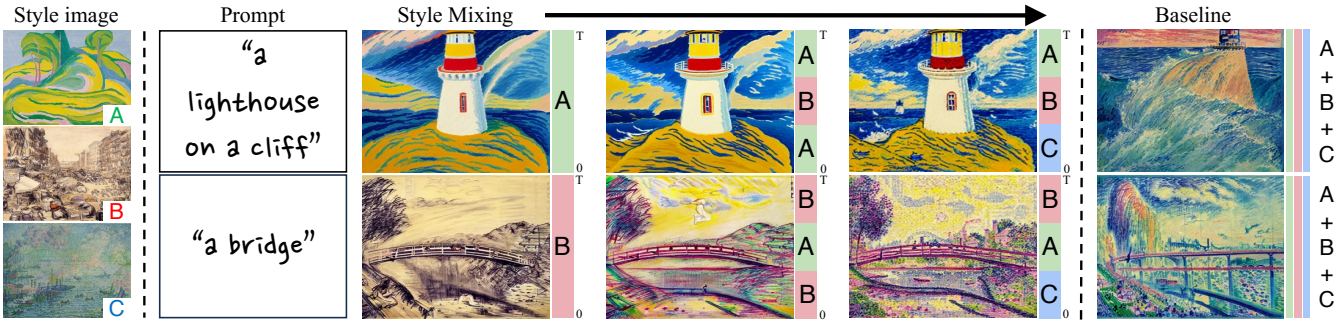

Figure 12: Style mixing. Multi-stage TI facilitates style mixing from various style references. A user can customize a new style by substituting style tokens at different stages t. For example, the style token closer to $t = T$ tends to influence the structure of the image, while those closer to $t = 0$ have a stronger effect on local and detailed attributes. For comparison, we display the baseline that employs all style tokens at every stage (*i.e.* using "A painting in S_t^A, S_t^B, S_t^C style" at all stages).

tically congruent, might diverge from the intended context. On the other hand, strengthening context guidance (λ_c) ensures the output resembles the inference prompt, highlighting context guidance's essential role in preserving contextual integrity. However, excessively strong context guidance could steer the output away from the original style, underlining the need for a nuanced balance of guidance for generating visually appealing and contextually accurate images. Nevertheless, this offers a new dimension of control over the synthesized image, differing from the classifier-free guidance (Ho and Salimans 2022). The additional control is a crucial element in the workflow of digital art production, considering its delicate and nuanced final outcomes.

Style mixing. As shown in Figure 12, multi-stage TI opens up a novel avenue for an intriguing aspect of style mixing from diverse style references. This process empowers users to customize a unique style by deploying different style tokens at each stage t. The style tokens close to $t = T$ predominantly impact the structure of the image, akin to broad strokes, while tokens closer to $t = 0$ affect local and detailed attributes, akin to intricate brushwork. To provide a concrete point of comparison, we present a baseline model that incorporates all style tokens at every stage, using the prompt "A painting in S_t^A, S_t^B, S_t^C styles". While the baseline produces reasonable style quality, it lacks a control factor for extracting partial stylistic features from the reference. Consequently, the fusion of styles with multi-stage TI underscores the creative and flexible nature of our model, offering users a broad range of applications for artistic creation.

Conclusion

We have introduced DreamStyler, a novel image generation method with a given style reference. By optimizing multi-stage TI with a context-aware text prompt, DreamStyler achieves remarkable performance in both text-to-image synthesis and style transfer. Content and style guidance provides a more adaptable way of handling diverse style references.

Limitations. While DreamStyler exhibits outstanding ability in generating artistic imagery, it is important to acknowledge its limitations within the intricate context of artistic expression. The vast spectrum of artistry, spanning from primitive elements to more nuanced and abstract styles (such as surrealism), demands thorough definition and examination from both artistic and technological perspectives.

References

Ahn, N.; Kwon, P.; Back, J.; Hong, K.; and Kim, S. 2023. Interactive Cartoonization with Controllable Perceptual Factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16827–16835.

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; Catanzaro, B.; et al. 2022. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

Chen, H.; Wang, Z.; Zhang, H.; Zuo, Z.; Li, A.; Xing, W.; Lu, D.; et al. 2021. Artistic style transfer with internal-external learning and contrastive learning. *Advances in Neural Information Processing Systems*, 34: 26561–26573.

Choi, J.; Lee, J.; Shin, C.; Kim, S.; Kim, H.; and Yoon, S. 2022. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11472–11481.

Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Pan, X.; Wang, L.; and Xu, C. 2022. StyTr2: Image Style Transfer with Transformers. In *CVPR*, 11326–11336.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *CVPR*, 2414–2423.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hong, K.; Jeon, S.; Lee, J.; Ahn, N.; Kim, K.; Lee, P.; Kim, D.; Uh, Y.; and Byun, H. 2023. AesPA-Net: Aesthetic Pattern-Aware Style Transfer Networks.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 1501–1510.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, W.; Xu, X.; Xiao, X.; Liu, J.; Yang, H.; Li, G.; Wang, Z.; Feng, Z.; She, Q.; Lyu, Y.; et al. 2022. UPainting: Unified Text-to-Image Diffusion Generation with Cross-modal Guidance. *arXiv preprint arXiv:2210.16031*.

Liu, S.; Lin, T.; He, D.; Li, F.; Wang, M.; Li, X.; Sun, Z.; Li, Q.; and Ding, E. 2021. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6649–6658.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Sohn, K.; Ruiz, N.; Lee, K.; Chin, D. C.; Blok, I.; Chang, H.; Barber, J.; Jiang, L.; Entis, G.; Li, Y.; et al. 2023. StyleDrop: Text-to-Image Generation in Any Style. *arXiv preprint arXiv:2306.00983*.

Tan, W. R.; Chan, C. S.; Aguirre, H.; and Tanaka, K. 2019. Improved ArtGAN for Conditional Synthesis of Natural Image and Artwork. *IEEE Transactions on Image Processing*, 28(1): 394–409.

Voynov, A.; Chu, Q.; Cohen-Or, D.; and Aberman, K. 2023. $P+$: Extended Textual Conditioning in Text-to-Image Generation. *arXiv preprint arXiv:2303.09522*.

Wang, Z.; Zhang, Z.; Zhao, L.; Zuo, Z.; Li, A.; Xing, W.; and Lu, D. 2022. AesUST: towards aesthetic-enhanced universal style transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1095–1106.

Zhang, L.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.

Zhang, Y.; Huang, N.; Tang, F.; Huang, H.; Ma, C.; Dong, W.; and Xu, C. 2023. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10146–10156.