StyleBoost: A Study of Personalizing Text-to-Image Generation in Any Style using DreamBooth

Junseo Park, Beomseok Ko and Hyeryung Jang
Department of Artificial Intelligence
Dongguk University
Seoul, South Korea
emails: {mki730, roy7001, hyeryung.jang}@dgu.ac.kr

***Abstract*—Recent advancements in text-to-image models, such as Stable Diffusion, have demonstrated their ability to synthesize visual images through natural language prompts. One approach of personalizing text-to-image models, exemplified by DreamBooth, fine-tunes the pre-trained model by binding unique text identifiers with a few images of a specific subject. Although existing fine-tuning methods have demonstrated competence in rendering images asking to the styles of famous painters, it is still challenging to learn to produce images encapsulating distinct *art styles* due to abstract and broad visual perceptions of stylistic attributes such as lines, shapes, textures, and colors. In this paper, we present a new fine-tuning method, called StyleBoost, that equips pre-trained text-to-image models to produce diverse images in specified styles from text prompts. By leveraging around 15 to 20 images of StyleRef and Aux images each, our approach establishes a foundational binding of a unique token identifier with a broad realm of the target style, where the Aux images is carefully selected to strengthen the binding. This dual-binding strategy grasps the essential concept of art styles and accelerates learning of diverse and comprehensive attributes of the target style. Experimental evaluation conducted on three distinct styles - realism art, SureB art, and anime - demonstrates substantial improvements in both the quality of generated images and the perceptual fidelity metrics, such as FID and CLIP scores.**

***Index Terms*—text-to-image models, diffusion models, personalization, fine-tuning**

I. Introduction

The domain of text-to-image generation has shown remarkable advancements [1]–[5] fueled by the emergence of sophisticated models such as Stable Diffusion [5]. These models exhibit an impressive ability to synthesize intricate visual representations from text inputs, enabling the generation of diverse images through natural language prompts. Recently, the pursuit of personalizing text-to-image models has gained considerable attention, aiming to empower diffusion-based text-to-image models with the capability to infuse personalized attributes of interest, such as a user-provided object, into their generated outputs. A remarkable approach in this direction is DreamBooth [6], a fine-tuning technique of a pre-trained text-to-image model to combine a unique text identifier with a limited set of images associated with a specific object. This method has enhanced the model's adaptability to individual preferences, enabling producing images of the subject in diverse scenes, poses, and views.

In the area of personalizing text-to-image synthesis, while significant progress has been achieved in synthesizing images mimicking the art styles of renowned painters and artistic icons, such as Van Gogh or Pop Art, it is still challenging to learn to generate images encapsulating a broader spectrum of artistic styles. The concept of an "art style" encompasses an intricate fusion of visual elements such as lines, shapes, textures, and spatial and chromatic relationships, and landscapes and/or subjects representing a specific art style are lied in a wide range, e.g., "`an Asian girl in the style of Van Gogh`", "`a London street in the style of Van Gogh`". Therefore, in contrast to specific objects, personalizing text-to-image synthesis capable of generating images aligned with artistic styles needs to convey abstract and nuanced visual attributes that are rarely classified or quantified. Due to these wide characteristics of images in art styles, we observe that applying existing personalizing methods, such as DreamBooth [6] or Textual Inversion [7], to mimic the art styles is not straightforward enough to yield good performance.

In this paper, we address this challenge by introducing a novel fine-tuning method called *StyleBoost*. Our approach leverages the capabilities of pre-trained text-to-image models, enabling them to produce a rich spectrum of images that adhere to specific artistic styles, or a target style, with the guidance of text prompts. Using 15 to 20 images that exemplify the distinctive characteristics of the target style, along with auxiliary images, StyleBoost seeks to capture the intricate and detailed aspects of the target style. In particular, it involves a dual-binding strategy: first, establishing a foundational connection between a unique token identifier, e.g., "`A [V] style`", and the general and broad features of the target style; and second, employing auxiliary images with the auxiliary token, e.g., "`A style`", to embed general aspects of the artwork, especially essential information of creating a person, and further to boost the acquisition of diverse and comprehensive attributes inherent to the target style. We measured FID for evaluating whether the three styles are similar to the original image. In addition, the performance evaluation was conducted through the CLIP score experiment of evaluating text image alignment. The results all scored better than the DreamBooth model for each style. Our main results are illustrated in Fig. 1.

II. Related Works

Prior research has delved into diffusion-based text-to-image generation methods, and personalization of text-to-image syn-

Fig. 1: Text-to-image synthesis of StyleBoost. Personalized images generated by StyleBoost compared to the existing DreamBooth for 3 different styles. Across the categories of person, animal, and background (landscape), our model generates meticulously aligned, high-fidelity images that resonate with the target style.

thesis. Early text-based image generation models, like GANs [8], struggled with representing images with diverse subjects and worked primarily with well-structured objects (e.g., faces and automobiles). The advent of advanced diffusion models and transformers yielded models like DALL-E [1], DALL-E2 [2], Imagen [3], Stable Diffusion [5], and ControlNet [9]. Trained on extensive text-image pairs, these models demonstrated impressive results in generating complex scene's structure and semantics. However, consistent preservation of a particular subject's identity across synthesized images remained challenging.

Efforts like Textual Inversion [7] and DreamBooth [6] addressed this gap by leveraging pre-trained text-to-image diffusion models for personalized image synthesis. Towards a similar goal of our approach, DreamBooth and Textual Inversion aim to extend the output domain of image synthesis, adopting a specific concept or identity given as input. These methods employ transfer learning by fine-tuning parameters within a text-to-image diffusion model, optimizing word vectors for new concepts. Notably, a recent work [10] enhanced personalization efficiency by selectively modifying the Attention layer of U-Net, akin to DreamBooth. Our approach builds upon DreamBooth's foundation to enhance personalized synthesis focusing on various art styles by introducing auxiliary and refined image composition.

III. METHOD

A. Preliminary: DreamBooth

Diffusion models [11] are probabilistic generative models that learn a data distribution by gradual denoising an initial Gaussian noise variable. Diffusion-based text-to-image generation, including Imagen, Stable diffusion, and DALL-E, combines the language models with diffusion models to encode capture meaningful representations of text descriptions, which is effective for the text-to-image synthesis. DreamBooth [6] is a recent method for personalization of pre-trained text-to-image diffusion model, such as Stable Diffusion, using only a few images of a specific object, called *instance images*. Using $3-5$ images of the specific object (e.g., my dog) paired with a text prompt (e.g., "`A [V] dog`") consisting of a unique token identifier (e.g., "`[V]`") representing the given object (i.e., my dog) and the corresponding class name (e.g., "`dog`"), DreamBooth fine-tunes a text-to-image diffusion model to encode the unique token with the subject.

To this end, DreamBooth introduces a class-specific prior preservation loss that encourages the fine-tuned model to keep semantic knowledge about the class prior (i.e., "`dog`") and produce diverse instances of the class (e.g., various dogs). Specifically, denoting a pre-trained text-to-image diffusion model by $\hat{\mathbf{x}}_\theta$, with θ being model parameters, it generates an image $\mathbf{x} = \hat{\mathbf{x}}_\theta(\boldsymbol{\epsilon}, \mathbf{c})$ given an initial noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a conditioning vector $\mathbf{c} = \Gamma(\mathbf{P})$ obtained from a text encoder Γ and a text prompt $\mathbf{P}$. For prior preservation, images of the class prior $\mathbf{x}_{\text{pr}} = \hat{\mathbf{x}}(\boldsymbol{\epsilon}_{\text{pr}}, \mathbf{c}_{\text{pr}})$ are sampled from the frozen pre-trained model $\hat{\mathbf{x}}$ with noise $\boldsymbol{\epsilon}_{\text{pr}}$ and conditioning vector $\mathbf{c}_{\text{pr}}$ corresponding to the class name (i.e., "`dog`"); and the model $\hat{\mathbf{x}}_\theta$ is fine-tuned using reconstruction loss for both instance images $\mathbf{x}$ of the specific object and class prior images $\mathbf{x}_{\text{pr}}$ to successfully denoise latent codes $\alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$ over the diffusion process t. The loss is written as follows (see [6] for the details):

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}\Big[w_t \|\hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x}\|_2^2 + \lambda w_{t'} \|\hat{\mathbf{x}}_\theta(\alpha_{t'} \mathbf{x}_{\text{pr}} + \sigma_{t'} \boldsymbol{\epsilon}', \mathbf{c}_{\text{pr}}) - \mathbf{x}_{\text{pr}}\|_2^2\Big], \quad (1)$$

where λ controls the relative weight of the prior-preservation term. In [6], DreamBooth is shown to synthesize depictions of a subject with a guidance of a text prompt by embedding the given subject to a unique identifier by training via (1) with a few images of the subject.

B. StyleBoost: Overall Framework

Our goal is to fine-tune the pre-trained text-to-image diffusion model so that it can generate novel renditions of a specific *art style*, or simply a style, using the guidance of a text prompt. To this end, we adopt a framework of DreamBooth [6] but aim at learning to synthesize images of a specific style of interest, which we call *target style*, instead of focusing on a particular

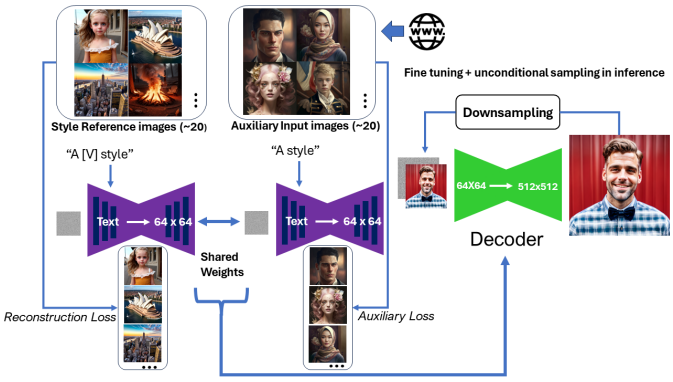


Fig. 2: The architecture of StyleBoost. StyleRef images of the target style, paired with text prompt ("`A [V] style`"), and Aux images, collected from the Internet and paired with the prompt ("`A style`") are provided as input images. After fine-tuning, the text-to-image model can generate various images of target style with guidance of text prompts.

object. Simply, our approach, dubbed *StyleBoost*, fine-tunes the text-to-image diffusion models $\hat{\mathbf{x}}_\theta$ by using a few reference images $\mathbf{x}$ of the target style, which we call **StyleRef images**, and a set of auxiliary images $\mathbf{x}_{\text{aux}}$, called **Aux images**, which is carefully selected to boost binding of the target style. The loss of StyleBoost is given as follows:

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}\Big[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\|_2^2 + \lambda w_{t'}\|\hat{\mathbf{x}}_\theta(\alpha_{t'}\mathbf{x}_{\text{aux}}+\sigma_{t'}\boldsymbol{\epsilon}',\mathbf{c}_{\text{aux}})-\mathbf{x}_{\text{aux}}\|_2^2\Big]. \quad (2)$$

We note that the overall training architecture follows the details of DreamBooth, as illustrated in Fig. 2. However, the novel point of our method is to investigate and analyze practical configurations of the StyleRef and Aux images, whose roles should be re-designed towards capturing broad characteristics of the target style.

As pointed out in [6], personalization of text-to-image models commonly leads to some issues of *(i)* overfitting to a small set of input images (i.e., StyleRef images), thus creating images of a particular context and subject appearance (e.g., pose, background), of breaking phenomenon, or lack of text-image alignment; and *(ii)* language drift that loses diverse meanings of the class name (i.e., "`style`" in our case) and causes the model to associate the prompt with the limited input images. Focusing on personalizing models to fit a style, not a subject (e.g., dog), we note that language drift no longer poses a significant concern since "`style`" is an abstract concept such that there is less need to keep the meaning and diversity of the word. In our findings, as shown in Sec. IV-D, a pre-trained text-to-image model (e.g., Stable Diffusion v1.5) embeds a broad set of images to the word 'style', i.e., from fashion styles, fabric pattern styles, to art styles. Rather than keeping unnecessary meanings of the word, we propose to use Aux images $\mathbf{x}_{\text{aux}}$ to allow the token of "`style`" to encapsulate the concepts such as *artwork* and/or *people* which are essential and useful to personalization in a target style.

As illustrated in Fig. 2, in StyleBoost, we provide $15-20$ images $\mathbf{x}$ of target style, paired with text prompts containing a unique token identifier (e.g., "`A [V] style`"), where the composition needs to be carefully designed to include comprehensive information of the target style, i.e., in aspects of background and people. Additional issues come with personalizing in a target style. In particular, style is an abstract concept that cannot be defined explicitly, thus understanding its contextual knowledge and subtle nuances, captured by relationships between various visual elements, is challenging; and especially, it is hard to draw people (e.g., face details) while keeping the style. To tackle these issues, around 20 Aux images collected from the Internet are provided, also paired with prompt "`A style`", with the purpose of preventing overfitting, boosting the StyleRef images to bind the target style, thus improving text-to-image synthesis performance. The according roles of Aux images will be outlined in the following subsection.

C. Rationale behind Auxiliary images

We recall that the advantage of DreamBooth [6] is to provide a means to normalize the entire model while learning it through class prior images. We transform this tool into the name of Aux images for the purpose of style personalization, where three main roles of the Aux images are discussed here.

Providing general embeddings of art styles. In most pre-trained models, the token 'style' primarily encapsulates the concept of fashion styles. In contrast, we need the token to be more related to 'artworks' and 'people' that help create images of a target style rather than a fashion style. We aim to bind the prompt for Aux images (i.e., "`A style`"), which is also a part of the prompt for StyleRef images (i.e., "`A [V] style`"), with general aspects of the artwork, especially providing essential information of creating a person. For this purpose, in StyleBoost, a set of Aux images includes diverse high-resolution images of a person drawn in any style (e.g., digital painting, Cubism). As will be seen through empirical results in the next section, it leads to a positive effect to improve the overall learning performance by re-directing embedding of the word 'style' from fashion style to general artwork-like styles, and to mitigate language drift and overfitting.

Aiding in the binding of the target style. Binding a unique token identifier with an object is relatively easy, however, we need to capture general and various features of the target style in broad aspects from the StyleRef images and further need to bind a unique identifier with them. Since the variation degree of any style is large, such as vibrant colors, exaggerated facial expressions, and dynamic movement in the case of 'anime' style, learning to bind the target style with only a few StyleRef images becomes more challenging. To boost the binding process, we leverage Aux images that are similar to the target style but with more general attributes, and use the according prompt to bind a part of the target style. As a result, StyleRef prompt ("`A [V] style`") captures more detailed components of the target style, while Aux prompt

("A style") contains general information that contributes to a comprehensive understanding of the target style. In other words, Aux images act as a secondary binding in personalizing the target style.

Improving text-to-image performance. In StyleBoost, a set of Aux images $\mathbf{x}_{\text{aux}}$ is, instead of being generated by a pre-trained diffusion model as in DreamBooth, collected as high-resolution images from the Internet with the purpose of improving the overall performance of text-to-image synthesis. Expressing detailed descriptions of a person (e.g., hands, legs, facial, and full-body shot) is still a critical point in qualitative evaluations, while drawing landscapes or animals are observed to be less sensitive. Therefore, our Aux images primarily consist of portraits and/or people, which we believe helps yield high-quality images for person-related prompts.

D. Extension : Auto-inpainting

Even though our approach StyleBoost is capable of synthesizing images of the target style from text prompts, it still has room to improve detailed facial descriptions. We propose a post-processing method of *auto-inpainting* that uses YOLOv5 [12], inpainting, and upscaling modules. After generating an image using a fine-tuned model, we first apply YOLOv5 for face detection, followed by upscaling the original image to obtain a 512×512 size of the bounding box to be masked. The cropped region, centered around the masking coordinates, is further refined to 256×256 using an inpainting model, which is then superimposed onto the original image. This process yields a more precise facial representation due to the upscaling and re-generation of the masked area and its surrounding pixels. Additionally, it facilitates the restoration of low-quality portions of facial descriptions.

IV. Experimental Results

In this section, we evaluate the performance of the proposed method StyleBoost on personalizing text-to-image generation to any style. We also provide comprehensive analysis of each individual component of StyleBoost.

A. Experimental setup

We conduct experiments on three common styles, namely realism art, surrealism and baroque art (SureB art), and anime styles, where the characteristics of each style are summarized as follows:

- Realism art is a style that emphasizes objective representation by depicting the subject accurately and in detail.
- SureB art is a style that blends pragmatic elements with dramatic, exaggerated expressions, i.e., a hybrid of surrealism and baroque art, creating the union of dreamlike visuals with real-world components.
- Anime refers to a Japanese animation style characterized by vibrant colors, exaggerated facial expressions, and dynamic movement.

Example images for each style are provided in the leftmost column of Fig. 1. We note that it can be applied to arbitrary style. For fine-tuning, we reference the target style's unique identifier using "a photo of [V] style", and the Aux prompt using "a photo of style". As a base diffusion model, we use stable diffusion (SD v1.5) model, which is pre-trained with realistic images.

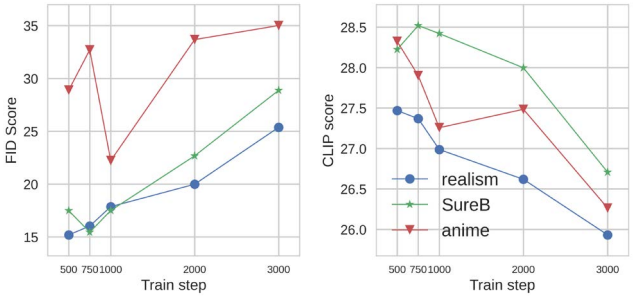
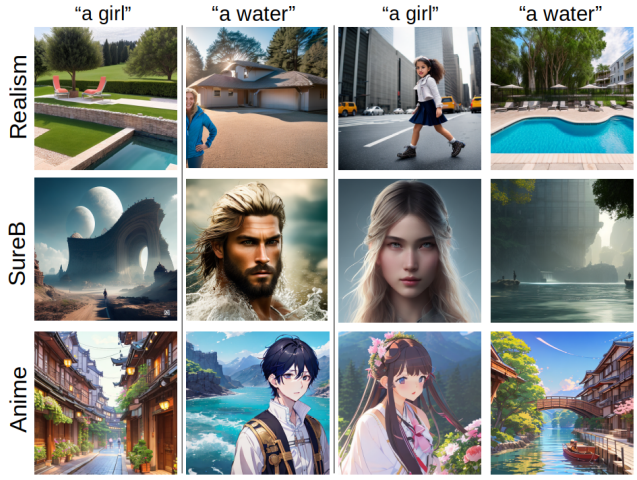


Fig. 3: FID and CLIP scores of generated images as a function of processed fine-tuning steps for different target styles. (left) Best FID score is achieved at 500, 750 and 1000 steps for realism art, SureB art and anime styles, respectively. (right) Text-to-image alignment is better with fewer training steps.

For evaluation, we use FID [13] and CLIP [14] scores to assess the quality of generated images. FID score quantifies the statistical similarity between real and generated images' visual features, where a lower FID score indicates more similarity. CLIP score measures the image-text alignment, where a higher CLIP score corresponds to being more aligned. We select $1,562$ prompts from Parti Prompts [4], e.g., "the Eiffel Tower", "a cat drinking a pint of beer", and "a scientist". The prompts used cover approximately 12 categories including people, animals, artwork, cars, and food, and we generate 12 images per prompt, total $18,744$ images of the target style. On the other hand, fake style that learned the target style generates 6 images per prompt, total $9,372$ images for each style. We collect images that effectively capture the essence of each target style using pre-trained diffusion models from Hugging Face [15].

B. Training configuration and style personalization

Our study begins with the examination of training configurations. We employed the Adam optimizer with a learning rate of $1e{-}6$, setting inference steps at 30 and a λ value of 1 for the experiments. Fine-tuning the pre-trained text-to-image model for three target styles involved minimizing the loss (2) using 20 StyleRef images and 20 Aux images. Fig. 3 reveals the necessity of adjusting training steps for each style to optimize text-to-image synthesis. The task of personalizing the base model (SD v1.5), which mostly generates realistic art images, with a target style of Realism proves relatively uncomplicated (best FID score at 500 steps). In contrast, adapting it for styles less aligned with the base model SD v1.5, such as SureB and Anime, requires extended iterations of 750 and 1000 steps, respectively. Fewer training steps mitigate overfitting, enhancing CLIP scores and text-image alignment. All ensuing experiments adhere to these training iterations. In other words, our approach that focuses on personalizing the style requires more training steps than object-based personalization



(a) Only back (b) Only person (c) Mixture of person and background

Fig. 4: Results of different compositions of StyleRef images. The prompt used for generation is written on the top of each column. Images in (a) and (b) show that the model understands the target style based on biased information, while generated images in (c) are well-aligned with the prompts.

methods, however, increasing training steps is not observed to enhance style representation universally since StyleRef images encompass a limited style range, neglecting many.

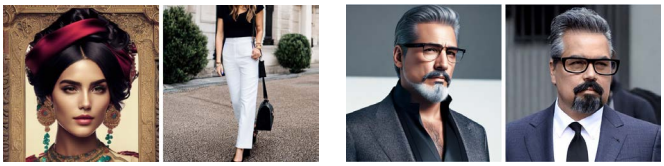
C. Analysis of StyleRef images

Here, we evaluate how StyleRef images impact personalization performance. Our broad style definition makes it challenging to encompass the target style using only 3-5 StyleRef images. If the variability of StyleRef images significantly increases within a limited number of training steps, learning can become difficult. Effectively learning from diverse StyleRef images with limited training steps requires managing their diversity. Experimental results show that approximately 20 images are needed for effective style learning.

We fine-tune SD v1.5 with various compositions of a set of 20 StyleRef images but without the use of Aux images: (i) 20 landscape images in the target style, (ii) 20 images of portraits and/or people in the target style, and (iii) a mixture of 10 landscape and 10 people images. When we use only landscape images for StyleRef images, the base model is observed to understand the target style based on spurious or biased information, e.g., the target style ("`a photo of [V] style`") indicates not to include any 'person'. Similarly, with only people images being used as StyleRef images, the appearance of 'person' becomes an essential element of "`a photo of [V] style`", resulting in overfitting. On the other hand, StyleRef images composed of 10 landscape and 10 people images can capture general features of the target style, including from landscape to person, thus successfully synthesizing images that are well-aligned with the prompts, as illustrated in Fig. 4.

D. Analysis of the Aux images

Now, we provide analysis of the Aux images. Given our utilization of a mere 20 images, we harness Aux images



(a) Aux prompt (b) Example reflecting Aux prompt

Fig. 5: Images are created by StyleBoost (left) and SD v1.5 (right), given prompts "`a photo of style`" (for (a)) and "`a photo of style, 45 year old man with goatee and gray hair, wearing elegant suit, wearing glasses`" (for (b)).

	FID (↓)			CLIP (↑)		
Method	Realism	SureB	Anime	Realism	SureB	Anime
w/o Aux	15.196	15.449	22.227	27.468	28.519	27.256
w/ Aux	**13.008**	**12.222**	**20.718**	**28.761**	**28.616**	**27.551**

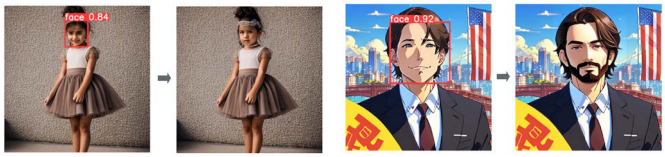
TABLE I: Auxiliary images ablation for three target styles, displaying FID and CLIP scores.

to augment performance. We composed the Aux image as a digital painting image. Digital painting offers the freedom to explore a wide range of styles, spanning from realism to abstraction. Leveraging digital tools, it attains a seamless and refined expression throughout. Moreover, by harnessing these tools to manipulate lines and colors with ease, it also grants the ability to convey fluid curves and shapes with utmost freedom. The commonality among the three styles we are experimenting with is their inclination towards digital painting, as they were generated using a diffusion model [11]. Therefore, we use a Aux image based on digital painting style to play a secondary binding role for the target style. Figure 5a and 5b show that the Aux prompt contains the style of digital painting. Among them, it contains information about drawing people. By moving the Aux prompt sharing the same token from the fashion style to the picture style area, you can see the direction that suits training purpose.

DreamBooth [6] mentions that using the Aux image can mitigate overfitting. In Table I, it can be seen that the FID performance improves overall when the Aux image is added. That is, the overfitting relaxation effect appears. In addition, since learning proceeds with fewer training steps, the mitigation effect is maximized. The effect of slightly increasing the CLIP score was also obtained. To sum up, Aux images are not directly related to the target style but can be used to complement the style learning process. These images contribute to a more comprehensive understanding of visual features and faithfully provide an auxiliary binding role for the target style.

E. Extension: Auto-inpainting

We propose an auto-inpainting method to alleviate the blurring of the face, which is a chronic problem of the text-to-image generation model. As an extension, this method shows superior performance in terms of qualitative evaluation rather



(a) Realism style (b) Anime style

Fig. 6: Auto-inpainting results for (a) realism and (b) anime styles. After face detection using YOLOv5 (left), inpainting is applied to reconstruct the face for the same prompt (right).

than quantitative evaluation such as FID, see Fig. 6. If the inpainting model is not the same as the target style, the FID is likely to drop. The inpainting model corresponding to each style used the model provided by Hugging Face [15].

F. Comparison

We compare our results with existing DreamBooth [6] method, using the hyperparameters provided in their work. Refer to Table II for the detailed comparison. We have confirmed that the existing DreamBooth method generally works normally, but its performance deteriorates for anime style. Utilizing the 'style' token for Aux image composition resulted in a realistic image that primarily revolved around the fashion style. In this setup, the generated Aux image lacked substantial information about the intended target style, and its directional alignment was also off the mark. On using the 'illustration Style' token, the generated image exhibited inconsistencies and struggled to fully encapsulate the essence of the target style. While it leaned towards resembling drawing styles, it still fell short in effectively complementing the target style. Furthermore, images produced using the frozen diffusion model displayed relatively lower quality compared to manually curated images, leading to potential performance degradation.

We also conducted experiments involving non-digital painting images as Aux inputs. Our image selection primarily aimed at those convertible to picture styles, while also conveying insights about human subjects. Notably, human-drawn art (e.g., Van Gogh) often carried a rugged and unnatural quality, failing to authentically encapsulate the nuances of the target style. Moreover, when actual target style images were integrated as Aux images, their contribution to comprehensive grasp diminished, manifesting instead as a direct influence that excessively heightened concentration on certain aspects. As a result, as demonstrated in Table II, our approach emerged superior to alternative methodologies.

V. Conclusion

We presented an approach that extends to abstract concept using a pre-trained text-to-image diffusion model for personalized image generation. The core idea is to transform the idea of DreamBooth [6] of regularization images. Extensive experimental results demonstrate that our approach uses fewer images while preserving text-image alignment due to fewer training steps. Ultimately, StyleBoost achieves remarkable style consistency at text-to-image synthesis.

Method	FID score (↓)	Realism art	SureB art	Anime
StyleRef images	Only back	22.804	24.598	34.629
	Only person	21.708	18.812	47.588
	Back + person	**15.196**	**15.449**	**22.227**
Aux images	Style token [6]	14.297	14.293	31.518
	Illustration style token [6]	14.093	14.466	28.570
	Human-drawn art	14.263	16.366	22.836
	Target style	15.855	13.990	29.450
	Digital painting (ours)	**13.008**	**12.222**	**20.718**

TABLE II: Comparison of FID performance for each style, with different compositions of StyleRef and Aux images. For Aux images, we chose StyleRef composition of Back+person.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2021R1F1A1063288), and under the ITRC (Information Technology Research Center) (IITP-2023-2020-0-01789) and the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00254592) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

[1] A. Ramesh, M. Pavlov *et al.*, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[2] A. Ramesh, P. Dhariwal *et al.*, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[3] C. Saharia, W. Chan *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022.

[4] J. Yu, Y. Xu *et al.*, "Scaling autoregressive models for content-rich text-to-image generation," *arXiv preprint arXiv:2206.10789*, vol. 2, no. 3, p. 5, 2022.

[5] R. Rombach, A. Blattmann *et al.*, "High-resolution image synthesis with latent diffusion models," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[6] N. Ruiz, Y. Li *et al.*, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 500–22 510.

[7] R. Gal, Y. Alaluf *et al.*, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022.

[8] S. Reed, Z. Akata *et al.*, "Generative adversarial text to image synthesis," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1060–1069.

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

[10] N. Kumari, B. Zhang *et al.*, "Multi-concept customization of text-to-image diffusion," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1931–1941.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[12] J. Redmon, S. Divvala *et al.*, "You only look once: Unified, real-time object detection," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[13] M. Heusel, H. Ramsauer *et al.*, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] J. Hessel, A. Holtzman *et al.*, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.

[15] "Hugging face," https:/huggingface.co.