Style Composition within Distinct LoRA modules for Traditional Art

Jaehyun Lee[1*] Wonhark Park[1*] Wonsik Shin[1] Hyunho Lee[1] Hyoung Min Na[2] Nojun Kwak[1]
[1]Seoul National University
[2]Kyunghee University
[1]{jhlee795, pwh0515, wonsikshin, hhlee822, nojunk}@snu.ac.kr, [2]leesan@khu.ac.kr

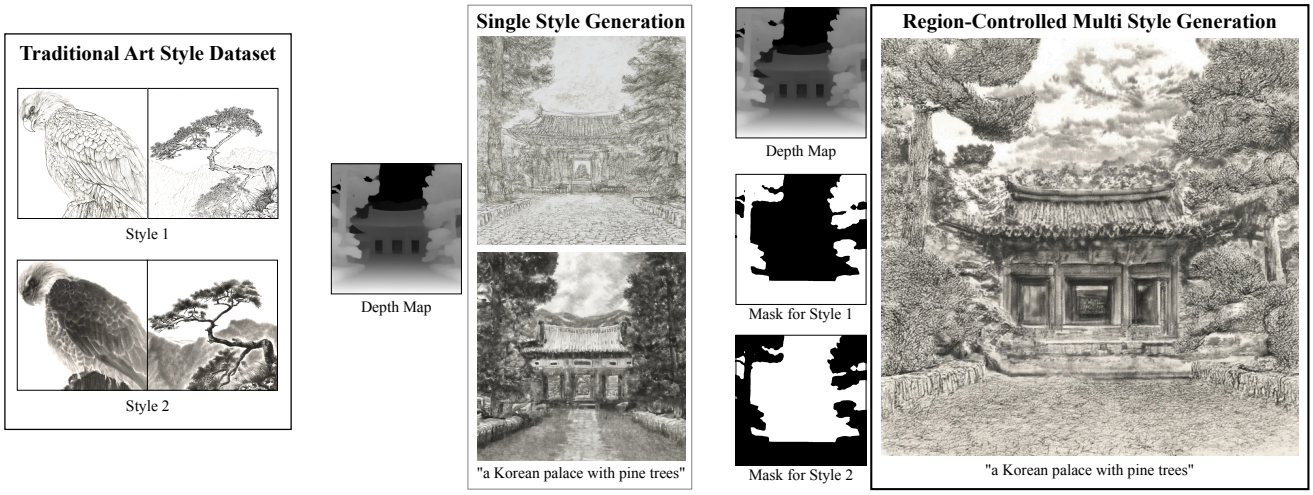

Figure 1. Given a style dataset, we train distinct LoRA modules for each individual style. Leveraging these specialized models, our method enables regional control over multiple styles within a single image, allowing users to generate art pieces that not only reflect their intended stylistic composition but also faithfully preserve the distinctive characteristics of each style.

Abstract

Diffusion-based text-to-image models have achieved remarkable results in synthesizing diverse images from text prompts and can capture specific artistic styles via style personalization. However, their entangled latent space and lack of smooth interpolation make it difficult to apply distinct painting techniques in a controlled, regional manner, often causing one style to dominate. To overcome this, we propose a zero-shot diffusion pipeline that naturally blends multiple styles by performing style composition on the denoised latents predicted during the flow-matching denoising process of separately trained, style-specialized models. We leverage the fact that lower-noise latents carry stronger stylistic information and fuse them across heterogeneous diffusion pipelines using spatial masks, enabling precise, region-specific style control. This mechanism preserves the fidelity of each individual style while allowing user-guided mixing. Furthermore, to ensure structural coherence across different models, we incorporate depth-map conditioning via ControlNet into the diffusion framework. Qualitative and quantitative experiments demonstrate that our method successfully achieves region-specific style mixing according to the given masks.

1. Introduction

Diffusion-based text-to-image (T2I) generative models [1, 2, 4–7, 23] have demonstrated impressive performance, enabling the creative generation of diverse images. Notably, through T2I style transfer [25, 29], these models can learn image representations with specific styles and generate desired outputs based on textual prompts. In the field of art, such models allow the learning of various artistic styles, facilitating not only creative expression but also the reproduction of traditional art aesthetics.

Despite the strong generative capabilities of T2I models, applying distinct artistic techniques to image generation remains a challenge. Artistic images often rely on clearly defined painting techniques, and multiple techniques are frequently blended in complex and user-intended ways. In par-

*Equal contribution.

ticular, traditional art requires adherence to specific styles while integrating multiple styles naturally. However, controlling natural blending of multiple styles is challenging due to entangled latent space in diffusion models [11, 27], unlike GANs [13, 14, 17, 18, 22, 28, 30]. As diffusion models do not support smooth interpolation in the latent space [8, 9, 20, 21], merging styles learned from separate models is difficult. A naive approach of linearly combining learned style embeddings often fails to provide controllable outputs; users cannot precisely apply specific techniques to the targeted regions, and one dominant style may overpower others. For artistic applications, it is essential to incorporate the artist's intention by allowing explicit control over where and how particular techniques are applied within the composition.

In this paper, we propose a zero-shot diffusion pipeline that enables natural style mixing, which refers applying different styles to user-specified regions (*e.g.* applying style 1 to the certain regions and style 2 to others) in a visually coherent way without blending the styles, by leveraging the strengths of separately trained diffusion models each specialized in a specific artistic style. We observe that latents at lower noise levels in diffusion contain stronger stylistic information, and that fusion is feasible even across different diffusion pipelines. Based on this insight, we introduce a method that performs style composition on the noise-clean latents predicted during the denoising process, enabling controllable multi-style synthesis. Also, for simultaneous generation with coherent structural content in various diffusion models, we attach ControlNet [31] to the diffusion model with depth map condition.

2. Related Work

2.1. Flow matching

Flow Matching [19] formulates generative modeling as learning a continuous-time ordinary differential equation (ODE) that transports samples from Gaussian noise to the data manifold by directly regressing a velocity field $v_\theta(x, t)$ along a prescribed interpolation path. Concretely, the model minimizes

$$\mathcal{L} = \mathbb{E}_{t\sim\mathcal{U}(0,1)} \left\| v_\theta\big(I_t(x_0, x_1),\, t\big) - \partial_t I_t(x_0, x_1) \right\|^2,$$

where $I_t(x_0, x_1) = (1-t)x_0 + t\, x_1$ linearly interpolates between noise x_0 and data x_1. At inference, given the trained velocity field v_θ, samples are obtained by numerically integrating the reverse-time ODE via an explicit Euler step:

$$z_{t_{k-1}} = z_{t_k} - \Delta t\, v_\theta(z_{t_k}, t_k),$$

where the step size $\Delta t = t_k - t_{k-1}$ is set by partitioning $[0, 1]$ into K steps, with $t_1 = 1$ and $t_0 = 0$. More recent work integrates Flow Matching into latent diffusion pipelines [1, 16] use a DiT architecture [21] to train their flow matching models, thereby enabling Flow-matching based high-quality image generation.

2.2. Style transfer

Style transfer re-renders a content image in the visual style of a reference image. It transfers color palettes, textures, and brushwork, while preserving the original scene's structure. The IP-Adapter [29] framework injects style via lightweight cross-attention adapters added in parallel to a frozen U-Net [24], capturing and transferring color, texture, and composition cues from a reference image with only 22 M extra parameters, and remains compatible with existing control tools. Its plug-and-play adapters enable efficient on-the-fly style adaptation without retraining the entire diffusion network, significantly reducing computational overhead. Recently, RB-Modulation [25] has enabled high-performance, optimization-based style transfer, dramatically reducing the required training time.

DreamBooth-LoRA [12, 26] specializes in style transfer by binding a dedicated style token to just a handful of reference images and using DreamBooth's prior-preservation loss to maintain the integrity of the original content. Instead of fine-tuning the entire model, it updates only a compact set of low-rank adapter weights, so that each module captures the essence of a particular visual style. During inference, you simply swap in the appropriate LoRA [12] module to transform any input subject into the target style, seamlessly transferring intricate artistic details while preserving the underlying scene structure. However, these methods were not inherently designed for the style-composition task of applying distinct styles to user-specified regions. Adapting them by simply masking styles onto the output image introduces unnatural artifacts and degrades overall quality. In contrast, our novel approach produces results that preserve natural coherence and visual fidelity.

3. Method

Unlike general images, traditional artworks often exhibit highly distinctive stylistic characteristics that vary across painting styles. Some artists aim to incorporate multiple style into a single painting by regions. This objective goes beyond conventional style mixing, which blends several styles to create a novel style. Instead, our style compositive sampling targets a more controlled generation process in which distinct styles are selectively applied to the designated regions within a single image. In this paper, we address the case compositing two distinct styles based on the given depth map. The overview of the entire method is provided in Figure 2.

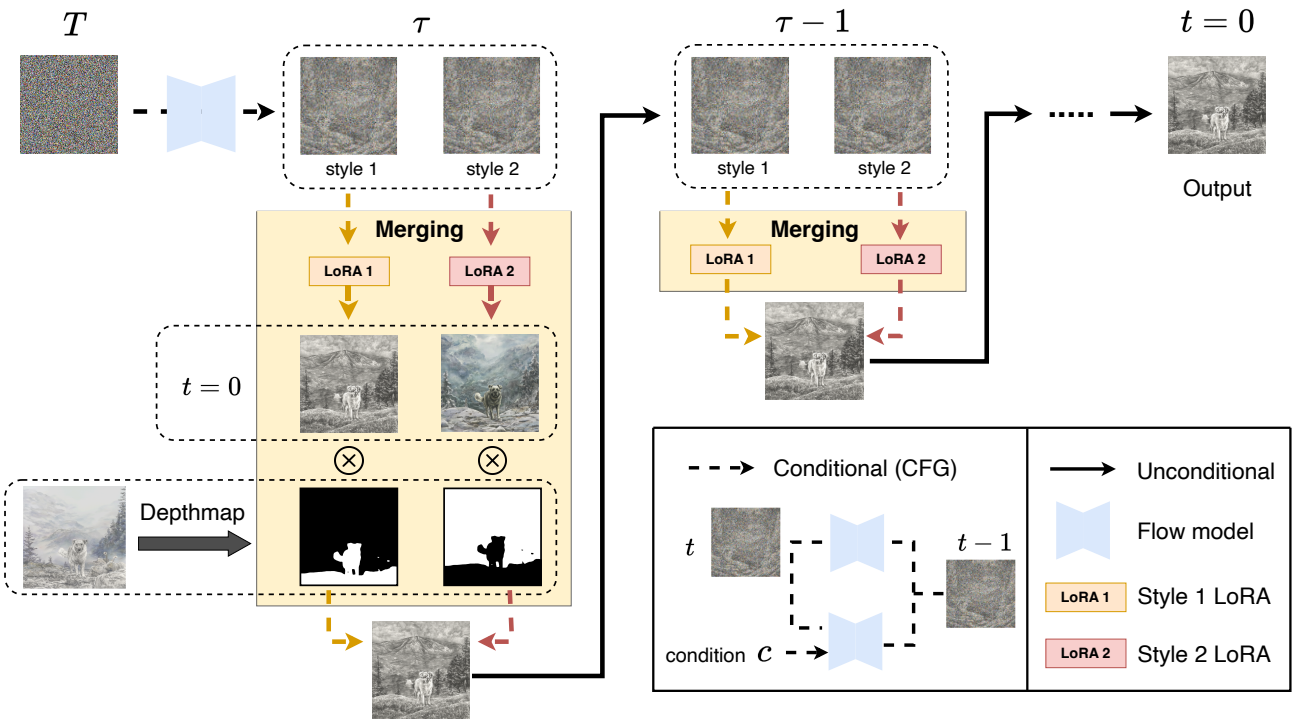

Figure 2. **Method overview.** Starting from a depth map D, two binary masks are extracted to define the target regions. Each mask guides a separate LoRA [12]-adapted diffusion pipeline (styles s_1 and s_2) through independent "simple updates" for coarse structure, followed by iterative "merging updates" that fuse per-style latents via Flow Matching scheduling and spatial masking. The result is a single image in which distinct styles are coherently applied to their designated regions.

3.1. Style LoRA Training

To enable per-style training, we employ Low-Rank Adaptation (LoRA) [12] modules, each dedicated to a single target style. Following DreamBooth [26], we introduce a rare placeholder token to indicate each style, thereby guiding the LoRA training of the diffusion model. Since the joint optimization of multiple styles within a single LoRA module significantly degrades generation quality, our framework adopts a one-style-per-LoRA module strategy. In fact, since our style compositive sampling enables compositing two distinct pipelines without any optimization of fine-tuning procedure, any style personalization methods which are able to be sampled through Flow Matching [19] can be integrated seamlessly to our framework.

3.2. Style Compositive Sampling

Given depth map D, we aim to generate an image in which specific regions are rendered in style s_1, while the remaining areas are rendered in style s_2, thereby achieving a coherent blend of two styles. To enable this, we extract two binary masks from the depth map D by thresholding and use them to generate each style in its designated region, which are used to fuse the latents during inference. Empirically, we observe that, as timesteps progress, the attention mechanism in model induces interactions between the regions in the latent space. Consequently, employing soft masks causes style-mixing issue, whereas binary masks yield the higher-quality, more coherent outputs. We consider two diffusion models, $v_\theta^{s_1}$ and $v_\theta^{s_2}$ each equipped with a LoRA module [12] trained to generate images in distinct styles s_1 and s_2, respectively. To incorporate structural information from depth map, we utilize ControlNet [31], conditioned on a given depth map D. Specifically, we employ the DiT model [21] combined with ControlNet, and we also adopts a Flow Matching Scheduling [19] where a pure noise sample at time $t = 1$ is denoised into an image at $t = 0$ by solving ordinary differential equation (ODE). In practical, these timesteps are used discretely from $t = T$ $(t = 1)$ to $t = 0$. For simplicity, we omit the notation for ControlNet in the followings.

The image generation process is divided into two levels, **simple updates** and **merging updates**. For the timesteps $t = T$ to $t = t'$, we apply **simple updates** that iteratively denoise each style's noise until a coarse layout becomes apparent. We delay the merge until the latent representation has acquired sufficient structural cues, thereby mitigating excessive mixing that would otherwise compromise merge fidelity. Specifically, at $t = T$, we sample pure noise

and then, for each style-specific model, compute the corresponding predicted velocity to iteratively update the latent representation. This process is repeated until timestep t'. Note that these denoising iterations proceed completely independently for each style.

From timestep t', we perform **merging updates** by first estimating, for each style s_i and current latent $z_t^{s_i}$, the conditional velocity $v_{t,c}^{s_i} = v_\theta^{s_i}(z_t^{s_i}, c)$ using the style prompt c and the unconditional velocity $v_{t,uc}^{s_i} = v_\theta^{s_i}(z_t^{s_i}, \varnothing)$ using the null prompt $\varnothing$. We then compute the total velocity $v_t^{s_i}$ by using classifier-free guidance [10]. Leveraging the ODE formulation of Flow Matching, we then approximate the clean latent $\hat{z}_0^i$ from the current noisy latent $z_t^{s_i}$ and its velocity $v_t^{s_i}$. These per-style endpoints are fused via a spatial mask to yield the merged latent $\hat{z}_0$. To return to the timestep t, we propagate $\hat{z}_0$ through each unconditional velocity $v_{t,uc}^{s_i}$ for each model, which is computed ealier. Notably, since two styles co-exist in the merged latent $\hat{z}_0$, we use the unconditional velocity to avoid artifacts that would arise from conditioning on mismatched style prompts. Finally, we compute the velocity for timestep $t-1$ based on the original estimate $v_t^{s_i}$ and merge it again through mask. We iterate this process until the denoising schedule completes.

Although latent mixing at timestep 0 was also proposed in [15], our problem involves blending styles rather than personalizing objects. As a result, no single prompt can reference two personalized styles simultaneously, and because the mask covers a far larger region than an object personalization, the attention mechanism exerts a far more influence on the generated images. Therefore, we had to propose an alternative novel approach as mentioned above to solve style composition task. Further details can be found in Algorithm 1.

Algorithm 1 Style Composition via Clean Latent Fusion

Input:
i-th Style LoRA flow-based generative models $v_\theta^{s_i}$
Style-mixing timestep τ, noise schedule σ_t
Source prompt embedding c, guidance scale w
Mask for i-th style $\mathcal{M}^{s_i}$
Output: Final style-composed image I_{style}
Sample initial noise $z_T \sim \mathcal{N}(0, I)$
Initialize i-th style latents $z_T^{s_i} = z_T, \forall i$
for $t = T$ to τ **do**
  $v_t^{s_i} = v_\theta^{s_i}(z_t^{s_i}, \varnothing) + w \cdot (v_\theta^{s_i}(z_t^{s_i}, c) - v_\theta^{s_i}(z_t^{s_i}, \varnothing))$
  $z_t^{s_i} = z_t^{s_i} - (\sigma_t - \sigma_{t-1}) \cdot v_t^{s_i}$
end for
for $t = \tau - 1$ to 0 **do**
  $v_t^{s_i} = v_\theta^{s_i}(z_t^{s_i}, \varnothing) + w \cdot (v_\theta^{s_i}(z_t^{s_i}, c) - v_\theta^{s_i}(z_t^{s_i}, \varnothing))$
  $\hat{z}_0^{s_i} = z_t^{s_i} - \sigma_t \cdot v_t^{s_i}$
  $\hat{z}_0 = \sum_i \hat{z}_0^{s_i} \cdot \mathcal{M}^{s_i}$
  $z_t^{s_i} = \hat{z}_0 + \sigma_t \cdot v_\theta^{s_i}(z_T^{s_i}, \varnothing)$
  $z_t^{s_i} = z_t^{s_i} - (\sigma_t - \sigma_{t-1}) \cdot v^{s_i}$
end for
$z_0 = \sum_i z_0^{s_i} \cdot \mathcal{M}^{s_i}$
$I_{style} \leftarrow \text{Decode}(z_0)$

4. Experiment

4.1. Traditional Dataset

We utilize an image dataset composed of Korean traditional art drawing techniques. The dataset includes five distinct styles, each with fifty images, where three for ink painting and the other two for color painting. Examples of all five styles are presented in Figure 4.

The ink painting techniques consist of Baekmyo, Gureuk, and Molgol styles.

- Baekmyo refers to technique of fine line drawing. It is a monochrome ink technique that emphasizes clean, precise outlines using fine brushwork without any coloring or shading. It captures the essence of a subject through contour lines alone, often used to depict figures or objects with clarity and restraint.
- Gureuk involves outlining the subject in ink and then filling the interior with color, where in ink painting, gray color. It maintains structure clarity while enabling subtle expression through fillings, balancing the strength of line and the depth of hue.
- Molgol is a "boneless" technique where forms are rendered directly with ink washes, without any prior outlines. It produces soft, flowing shapes that emphasize spontaneity and the natural movement of the brush.

The color painting techniques include Ilpil and Gongpil styles.

- Ilpil literally means "one stroke," and refers to a painting style where an object or shape is expressed in a single, bold brushstroke. It emphasizes fluidity, rhythm, and the painter's intuition, often used in expressive or abstract works.
- Gongpil is a meticulous coloring technique characterized by delicate lines and carefully controlled brushstrokes. It prioritizes precision, often used for highly detailed subjects such as court paintings or botanical illustrations.

4.2. Implementation Details

We adopted the recently proposed Flux-dev.1 model [16], which integrates DiT [21] with Flow Matching [19]. For LoRA [12] adaptation, we used trained each style for 3500 steps, and set the classifier-free guidance scale to 3.5. During inference, we set the number of inference steps as 28, and initiate the **merging updates** at $t' = 8$; *i.e.* after the first 20 denoising iterations. We use threshold value as 0.5 for extracting binary masks from the given depth map in our experiments. We conduct all experiments with four style

Figure 3. **Qualitative results.** Each box represents a pair of styles. Given a depth map, we present the output of individual style LoRA models on the left, and the result of style mixing on the right. Our method applies each style to specific regions using separate spatial masks, enabling fine-grained control over where each style is applied. In contrast, LoRA Fusion simply averages the trained LoRA parameters, lacking spatial awareness. As a result, our approach better preserves the characteristics of each style within their designated regions while achieving more coherent and natural blending across the entire image.

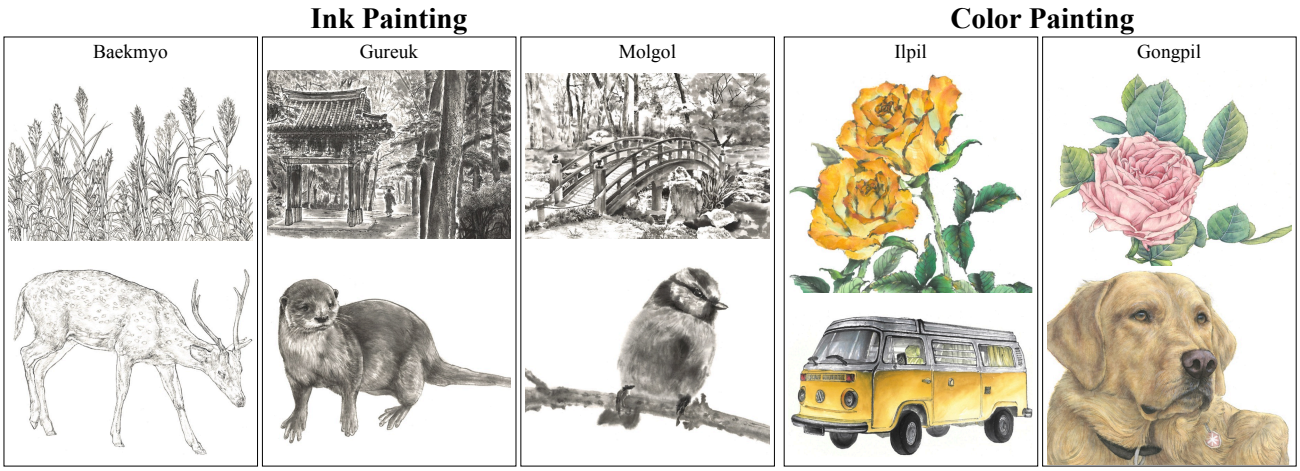

Figure 4. **Traditional Dataset.** Sample images from our Korean traditional art technique dataset, consisting of five representative styles. The dataset includes three ink painting techniques (Baekmyo, Gureuk, Molgol) and two color painting techniques (Ilpil, Gongpil). Each style exhibits unique attributes in line structure, stroke pressure of the brush, and color usage.

pairs, which are Baekmyo-Molgol, Baekmyo-Ilpil, Molgol-Gongpil, Ilpil-Gongpil.

Table 1. CLIP Image Scores between ground truth styles and the generated images

Style1	Style2	LoRA Fusion	**Ours**
0.673	0.676	0.692	**0.697**

4.3. Quantitative Experiments

We report quantitative evaluations for style 1, style 2, the naively weighted LoRA [12] blend(LoRA Fusion), and ours for each style pairs. To measure style similarity between generated and reference images, we employ the CLIP-I scores [3]. For both LoRA Fusion and ours, we apply spatial masks to isolate the regions for each style, compute CLIP-I scores on each masked region, and then average the two results. As reported in table 1, ours show better performance than simple LoRA Fusion. We conducted experiments on the four style pairs mentioned in sec 4.2. In each pair, style1 and style 2 denote the first and the second style, respectively. The reported score is the average of the scores obtained in these four pairs.

4.4. Qualitative Experiments

We also conduct qualitative experiments, comparing our method with a naive weighted blending of LoRA [12] modules (LoRA Fusion) which also mentioned in sec 4.3. Given the depth map and prompt presented in the first column, the style1 and style2 (second and third columns) illustrate the images generated by a model incorporating a single LoRA module trained on a single style for each case. Our method preserves the distinctive features of each style while allowing users to apply them to appropriate regions. To define style regions, we apply thresholding to the depth map and generate binary masks accordingly. In contrast, LoRA Fusion often suffers from style imbalance, where one style dominates the entire image or the blending process leads to the loss of distinctive characteristics associated with traditional art techniques. For instance, in 5th row (prompt "a dog in the living room" in Molgol - Gongpil case), ours constructs a coherent image in which Molgol's grayscale texture and Gongpil's color texture are distinctly presented in their respective masked regions; by contrast, in LoRA Fusion the grayscale texture dominates the entire image. Qualitative results are presented in Figure 3.

5. Conclusion and Limitation

We propose a zero-shot style compositive image generation pipeline that enables regional blending of multiple artistic styles using separately trained, style-specific LoRA modules [12]. By allowing users to explicitly assign styles to spatial regions via binary masks, our method offers fine-grained control over where and how each artistic technique is applied. Style composition is performed on the predicted noise-clean latents during each timestep of the flow matching denoising process, effectively capturing the detailed style information that each model possess. This approach not only preserves the fidelity of each individual style but also achieves smooth and coherent transitions across style boundaries. Furthermore, the modular design allows for open-ended extensibility, supporting the integration of arbitrary numbers of artistic styles without retraining a unified model.

A limitation of our current framework lies in its reliance

on ControlNet [31] with a depth map condition to maintain structural consistency across style-specific diffusion models. This dependency may limit flexibility in applications where accurate depth maps are not available.

Acknowledgements The researchers at Seoul National University were funded by the Korean Government through the grants from NRF (2021R1A2C3006659), IITP (RS-2021-II211343) and KOCCA (RS-2024-00398320).

References

[1] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 1, 2

[2] Alexander Quinn Nichol et al. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 16784–16804. PMLR, 2022. 1

[3] Alec Radford et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 6

[4] Chitwan Saharia et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022. 1

[5] Dustin Podell et al. SDXL: improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[6] Yogesh Balaji et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *CoRR*, abs/2211.01324, 2022.

[7] Shuyang et al. Gu. Vector quantized diffusion model for text-to-image synthesis. *arXiv preprint arXiv:2111.14822*, 2021. 1

[8] Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, and Humphrey Shi. Smooth diffusion: Crafting smooth latent spaces in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7548–7558, 2024. 2

[9] Jaehoon Hahm, Junho Lee, Sunghyun Kim, and Joonseok Lee. Isometric representation learning for disentangled latent space of diffusion models. In *International Conference on Machine Learning*, pages 17224–17245. PMLR, 2024. 2

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 4

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[12] 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. In *ICLR*, 2022. 2, 3, 4, 6

[13] Minguk et al. Kang. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[15] Gihyun Kwon and Jong Chul Ye. Tweediemix: Improving multi-concept fusion for diffusion-based image/video generation. In *ICLR*, 2025. 4

[16] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024. 2, 4

[17] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in neural information processing systems*, 32, 2019. 2

[18] Zhiheng Li, Martin Renqiang Min, Kai Li, and Chenliang Xu. Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[19] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. 2, 3, 4

[20] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36: 24129–24142, 2023. 2

[21] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2, 3, 4

[22] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1060–1069, New York, New York, USA, 2016. PMLR. 2

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. 1

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 2

[25] Litu Rout, Yujia Chen, Nataniel Ruiz, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. RB-modulation: Training-free stylization using reference-based modulation. In *ICLR*, 2025. 1, 2

[26] 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 *CVPR*, 2023. 2, 3

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2

[28] Tao et al. Xu. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2

[29] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. 1, 2

[30] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021. 2

[31] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2, 3, 7