ColorizeDiffusion: Improving Reference-based Sketch Colorization with Latent Diffusion Model

Dingkun Yan[1], Liang Yuan[2], Erwin Wu [1], Yuma Nishioka[1], Issei Fujishiro[2], Suguru Saito[1]
[1]Institute of Science Tokyo, School of Computing
[2]Keio University, Faculty of Science and Technology

Figure 1. Our method focuses on colorizing sketch images using reference images, especially anime-style images. By diminishing condition conflicts between input images, the proposed model can achieve visually pleasant results across a variety of contents and styles.

Abstract

Diffusion models have achieved great success in dual-conditioned image generation. However, they still face significant challenges in image-guided sketch colorization, where reference and sketch images usually exhibit different semantics and spatial structures. This mismatch, termed "distribution shift" in this paper, results in various artifacts and degrades the colorization quality. To address this issue, we conducted thorough investigations into the image-prompted latent diffusion model and developed a two-stage training framework to mitigate the effects of distribution shift based on our analysis. Comprehensive quantitative comparisons, qualitative evaluations, and user studies were performed to demonstrate the superiority of our proposed methods. Additionally, ablation studies were conducted to assess the impact of the distribution shift and the selection of reference embeddings. Codes are made publicly available at https://github.com/tellurion-kanata/colorizeDiffusion.

1. Introduction

Animation has been a popular artistic style worldwide for decades. The current workflow for creating anime-style images typically includes a line sketch followed by colorization, with the colorization step being especially labor-intensive and time-consuming. With the rapid development of deep learning, many effective network-based algorithms have been developed to automate this colorization process. Based on the type of guiding condition, existing methods can be categorized into three types: user-guided,

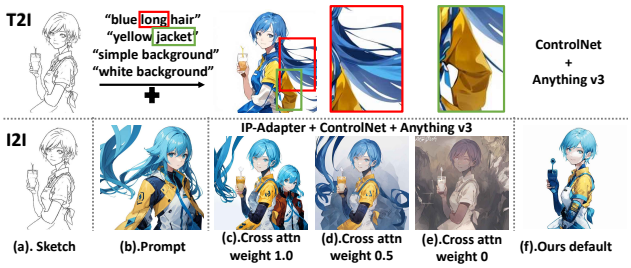

Figure 2. Illustration of spatial entanglement. The T2I model prioritizes prompt semantics and thus generates results with long hair and a jacket outside the sketch. Similar conflicts widely exist in I2I colorization but result in much worse artifacts, such as the extra person in column (c) and the messy background in column (d). Column (f) illustrates our result as correct colorization.

text-guided, and reference-based. Reference-based methods excel in colorizing images with a specific style and fine-grained textures, as both can be clearly expressed by the reference image. Yet, existing reference-based methods are ineffective in generating visually satisfying images in high resolution and are only applicable for limited inputs, such as figure-to-figure [30, 56, 63] or face-to-face [1, 3]. As these methods are developed based on generative adversarial nets (GANs) [13, 22] or small-scale diffusion models (DMs) [19, 50], they are incapable of transferring vivid strokes, textures, and backgrounds from references to sketches for arbitrary inputs.

Recently, text-to-image (T2I) DMs [41, 42] have achieved success on image generation tasks by enabling non-professional users to create artistic visual content with text prompts. Nevertheless, texts are limited as controlling signals for image generation as they struggle to convey precise spatial information for layouts, shapes, poses, textures, and styles. To address this limitation, researchers further proposed to combine image prompts into DMs: ControlNet and T2I-Adapter [35, 64] proposed to use spatial priors to guide DMs to generate images with identical layouts; IP-Adapter [57] integrated visual concepts extracted by CLIP [39] image encoders to pre-trained T2I models, enabling them for image-to-image (I2I) re-imagination.

However, image-prompted T2I DMs are still facing nontrivial problems on sketch colorization tasks in solving the potential conflicts between inputs during inference. Since sketches and reference images contain varied spatial embeddings about structure, layout, and segmentation, these semantic mismatches would likely cause visually unacceptable artifacts. We visualize this issue in Figure 2, where conflicting objects or identities are generated outside sketches. This specific type of artifact is called "spatial entanglement" in this paper.

To fully understand and diminish the impact of the semantic mismatches between reference images and sketches, we analyze this problem from the perspective of probability distribution within dual-conditioned training and name it "distribution shift" in this paper. Due to the absence of detailed embeddings in sketches and the semantic alignment between references and ground truth during training, the optimized distribution unavoidably shifts towards the prompt side and away from the sketch side. This deviation leads to a deterioration in image quality and a higher probability of generating artifacts during inference, especially for reference-based sketch colorization. To mitigate its negative impact, we propose a two-stage training scheme.

We designed a novel noisy training for the first stage, which adds timestep-dependent noise to the reference embeddings to reduce their information in the early denoising steps and, therefore, hinders the optimization of content/layout transfer. Narrowing the spatial information gap between reference images and sketches effectively reduced the spatial entanglement. A refinement stage follows the noisy training stage to enable fine-grained guidance on texture, strokes, and clear background. We conducted qualitative and quantitative comparisons and a user study with baselines to illustrate the superiority of the proposed method over existing methods. Ablation studies are also carried out to assess the impact of the distribution shift and the selection of reference embeddings, demonstrating the effectiveness of the proposed method in mitigating its negative impact.

To summarize, our contributions are as follows:

- We identify the distribution shift problem in image-prompted DMs for the sketch colorization task, which causes spatial entanglement and severely damages the quality of results.
- We design a two-stage training scheme for reference-based sketch colorization. The noisy training stage eliminates the effects of the distribution shift, and the refinement stage recovers the generation of fine strokes and clear backgrounds.

2. Related Work

Latent diffusion models. Diffusion Probabilistic Models [19, 50] are a class of latent variable models inspired by considerations from nonequilibrium thermodynamics [48]. Compared with Generative Adversarial Nets (GANs) [6,7,13,24,25], DMs excel at generating highly realistic images across various contexts. However, the autoregressive denoising process, typically computed using a U-Net network [43] or a Diffusion Transformer (DiT) [4, 37], incurs substantial computational costs. To address this limitation, Stable Diffusion (SD) [38, 42], a class of Latent Diffusion Models (LDMs), utilizes a two-stage synthesis and carries out the diffusion/denoising process within a highly compressed latent space within a pair of pre-trained Variational Autoencoder (VAE) to reduce computational costs significantly. Concurrently, many efficient samplers have been

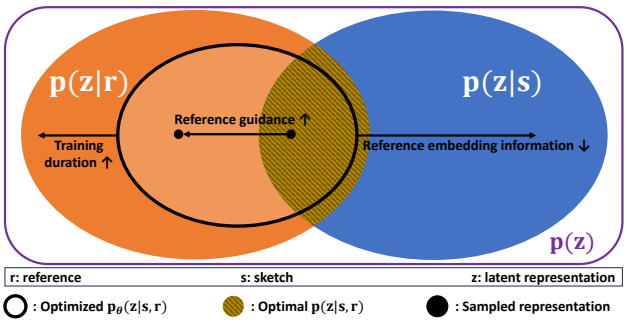

Figure 3. Illustration of the distribution shift. The optimized distribution gradually derivates from the optimal distribution, resulting in artifacts when reference images are semantically unaligned with sketches during inference. A solution is to reduce the information of reference embeddings during training.

proposed to accelerate the denoising process [32, 33, 49, 50]. We adopt SD as our neural backbone, utilize the DPM++ solver and Karras noise scheduler [23, 33, 50] as the default sampler, and employ classifier-free guidance [9, 20] to strengthen the reference-based performance.

Reference-based sketch colorization. Many effective methods have been developed to achieve automatic sketch colorization [11, 12, 17, 22, 36, 51, 65]. Depending on the type of guidance conditions, these methods can be classified into three types: reference-based [3, 30, 56], user-guided [61, 63], and text-guided [26, 56, 64]. Reference-based sketch colorization involves colorizing sketch images using reference images, allowing networks to generate specific strokes and textures, which are difficult to precisely describe using other conditions like text or palette. Existing reference-based methods differ in whether they adopt a pre-trained encoder to extract reference embeddings. Based on the experience of [41, 42, 56, 63], which have demonstrated superior performance in image quality and versatility, we adopt a pre-trained CLIP Vision Transformer (ViT) and freeze it when training the generative backbone.

Text/user-guided sketch colorization. Text-based [26, 64, 69] and user-guided methods [63, 66] have achieved impressive progress during the past years, driven by the rapid development of Large Language Models (LLMs) and T2I generative models. Alongside adapters such as ControlNet [28, 34, 62, 64] and T2I adapter [35, 52], T2I models can be applied to guided sketch colorization using various conditional inputs, including text and palettes. Moreover, IP-Adapter [57] introduces a copying mechanism to incorporate image prompts, allowing extended guidance with reference images. Therefore, reference-based colorization can also be achieved by jointly using ControlNet, IP-Adapter, and personalized T2I models. Such combinations achieve much better results than existing reference-based methods [3, 56] but suffer from the condition conflict discussed in this paper, which leads to much suboptimal results compared to the proposed model.

3. Methodology

3.1. Preliminary on diffusion and framework

DMs are generative models trained to learn a specific distribution from data by denoising variables sampled from a Gaussian distribution in T steps. To reduce the considerable computational cost, LDMs perform this denoising process within a perceptually compressed latent space encoded by a pre-trained VAE. The standard pipeline of training and inference is introduced in the supplementary materials.

The adopted neural backbone comprises a pre-trained VAE, a sketch encoder, a denoising U-Net θ, and a pre-trained Visition Transformer (ViT) from OpenCLIP-H [5, 21, 40, 46]. Given a reference image, the ViT outputs 1 CLS token and 256 local tokens. We adopt the local tokens as reference embeddings and inject them into the U-Net through cross-attention layers. The architecture of the U-Net is similar to SD v2.1 [42] and is initialized using Waifu Diffusion v1.4 [15].

We denote sketch images, reference images, and ground truth as s, r, and y, respectively. The pre-trained encoder, decoder, and U-Net are represented by $\mathcal{E}$, $\mathcal{D}$, and θ, respectively. The timestep t starts from $T-1$ and goes to 0, where T is the total number of diffusion steps, set to 1000. The ViT and extracted tokens are denoted as ϕ and τ_ϕ.

3.2. Distribution shift

Unlike text-guided or user-guided colorization, where user-given prompts exclude detailed spatial information and always correspond to the sketches semantically in training and inference, image-guided methods often involve spatial and semantic conflicts between inference inputs but are trained by fully matched pairs due to the difficulty of collecting presentable data. This gap between training inputs and inference inputs leads to severe deterioration in visual outcomes by degrading image quality and synthesizing numerous incompatible contents.

Given $p(z|y)$, the ground truth distribution, and $p(z|s)$ and $p(z|r)$, two ideal conditional distributions. Since the $p(z|r)$ always aligns with $p(z|s)$ and $p(z|y)$ during training, the optimized $p_\theta(z|s,r)$ deviates from $p(z|s)$ and moves towards $p(z|r)$ continuously as references contain much more semantics than sketches. While in the inference stage, $p(z|r)$ is usually out of $p(z|s)$. This gap results in visually unacceptable artifacts since the sampled features from $p_\theta(z|s,r)$ are more likely to be out of $p(z|s)$, as visaulized in Figure 3.

Moreover, image embeddings implicitly express size- and layout-related embeddings [38], which are likely to degrade the perceptual quality of reference-based results if accurately transferred to incompatible sketches. Unfor-

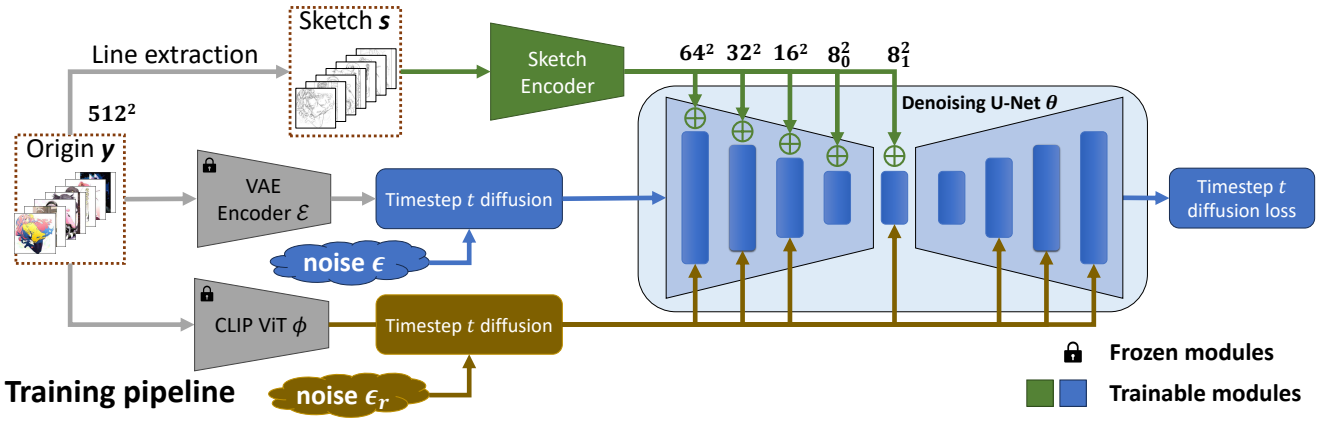

Figure 4. The architecture of our colorization model and the pipeline of the proposed noisy training. The noisy training is designed to optimize the style transfer and hinder the optimization of content transfer by adding timestep-dependent noise to image embeddings during training. Note that the noise added to the reference embeddings is not an optimization target.

tunately, vanilla reference-based training strengthens the transfer of these embeddings as the optimization progresses. This deterioration will be demonstrated by quantitative comparison among ablation models at different epochs.

A trade-off is adjusting hyperparameters to reduce the influence of image prompts, as shown in Figure 2. However, such adjustments are mostly ineffective for spatial entanglement and degrade the style similarity. Finding optimal combinations of hyperparameters for each input pair is also difficult and time-consuming.

3.3. Two-stage training

According to Figure 3, we mitigate the negative impact of the distribution shift by adding timestep-dependent noise to the reference embeddings, which reduces their information in early-step denoising and misaligning their semantics with ground truth. This method, termed noisy training, extends the training to improve the style transfer performance without causing severe deterioration in perceptual quality.

Noisy training. As previously analyzed in Section 3.2, longer training leads to a higher probability of generating incompatible content as the optimized distribution shifts away from $p(z|s)$ as training progresses. Many studies [10, 35, 67] have demonstrated that early sampling steps determine spatial semantics by rendering low-level features, such as layout and content, and subsequent steps refine these spots into detailed objects, identities, and textures. Inspired by these findings, we add timestep-dependent noise to reference embeddings during training to prevent content and layout transfer from being over-optimized. The added noise shrinks as timestep decreases from $T-1$ to 0, so we utilize the diffusion noise schedule directly. The pipeline of noisy training is illustrated in Figure 4.

Given α_t, β_t the hyperparameters of the noise scheduler at timestep t, the objective function of the proposed noisy training is formulated as:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathcal{E}(y),\epsilon,t,s,r}[\|\epsilon - \epsilon_\theta(z_t, t, s, \tau_{\phi,t}(r))\|_2^2], \quad (1)$$

where $\tau_{\phi,t}(r) = \alpha_t \tau_\phi(r) + \beta_t \epsilon_r$ and $\epsilon_r \sim \mathcal{N}(0,1)$. As the reference images used during training are ground truth, r can be replaced by y in this equation. We trained the network for five epochs and dropped 10% reference inputs for classifier-free guidance (CFG) [20].

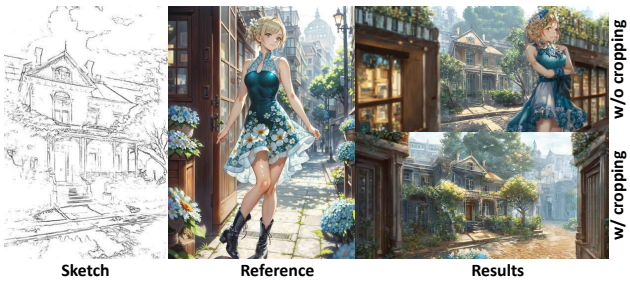

Figure 5. A comparison of inpainting. The upper result is generated by an ablation model trained without center cropping.

Refinement training. The optimized model cannot accurately generate fine-grained textures after the noisy training. To further improve synthesis quality, we added a refinement stage to recover the early-step layout/content transfer slightly after the network effectively transfers style-related embeddings. This fine-tuning follows the vanilla diffusion training, but 50% of reference inputs were dropped to avoid the distribution shift and 10% sketch inputs for CFG. We empirically set this fine-tuning to two epochs to avoid severe spatial entanglement. Results without second-stage training are included in the supplementary materials.

4. Experiment

For simplicity, we denote the reference-based CFG scale [20] as 'GS' and the cross-attention scale as 'CAS' in this

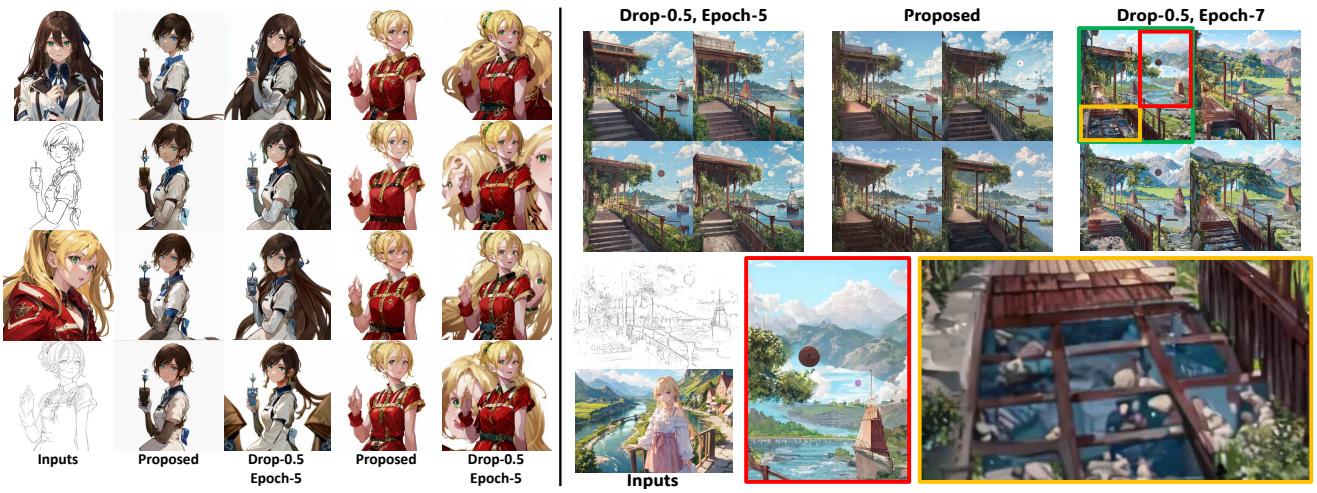

Figure 6. Results generated in one batch by respective models. As seen in the left comparison, the five-epoch *Drop-0.5* model shows a much higher probability of generating spatial entanglement compared to the proposed model. This tendency increases as training continues, highlighted in the right comparison, where compositions of results generated by the seven-epoch *Drop-0.5* model are visually chaotic.

section. To quantitatively estimate the perceptual quality of generated images, we utilize the Fréchet Inception Distance (FID) [18, 47], which quantifies the distance between the distributions of generated results and ground truth.

4.1. Implementation details

Dataset. We used Danbooru 2021 [8] as the original dataset to produce sketch images by jointly using SketchKeras [59] and Anime2Sketch [55]. The training set includes 4.8M+ pairs of (sketch, ground truth color) images. All quantitative evaluations were taken on a validation set, including 52,000+ ground truth tags and (sketch, color) image pairs. Samples of the training data are included in the supplementary materials.

Training and testing. We trained all models on an NVIDIA DGX-Station A100 with 4x NVIDIA A100-SXM 40G using Distributed Data-Parallel Training (DDP), DeepSpeed ZeRO2, and AdamW optimizer [27, 31], with a static learning rate as 0.00001. Our default sampling step for testing was set to 20. For all reference-based FID evaluations, the colorization was guided by randomly selected references.

Center cropping. Image-guided networks trained using both conditions show an inability to inpainting, caused by their sensitivity to sketch edges and view-related embeddings, which are implicitly expressed by image prompts. An example is shown in Figure 5. The upper result is semantically correct but visually unsatisfying due to its narrow composition. Consequently, we applied center cropping only to sketch inputs, while other pre-processings were simultaneously taken on both sketch and ground truth during training. Thus, the network learned to generate perceptually pleasant content in the margins. The effectiveness of center cropping degrades without the proposed noisy training.

Table 1. Quantitative comparison of FIDs with ablation models. We use uniform noise scheduler [50] for validation. Tested CFG scales are represented by GS-3 and GS-5, where optimal results are usually achieved. †: Tested at epoch 5. ‡: Tested at epoch 7.

Fréchet inception distance (50K-FID) ↓		
Model	GS-3	GS-5
CLS token, Proj-0.1	10.5273	10.3981
CLS token, CLS-0.1	17.6103	24.2609
†*Drop-0.5*	7.9077	8.2407
‡*Drop-0.5*	8.1842	9.1032
†*Noisy trained*	9.4761	10.9010
‡*Proposed model*	7.3676	**6.8551**

Specifically, image deformation has been widely used to produce reference images for training in previous methods [3, 29, 30, 56]. Yet, we found it degrades the quality of results without notably improving the spatial entanglement in experiments. Therefore, we discard this augmentation.

4.2. Ablation study and discussion

Architecture. We trained three important ablation models here to investigate the distribution shift, which we consider the primary cause of deterioration in reference-based sketch colorization. Its impact on dual-conditioned training is mainly manifested in a higher probability of spatial entanglement and degradation in composition and texture quality. All ablation models were trained for seven epochs as the proposed one.

1. Dropping model: This ablation model utilizes the same architecture as the proposed one but was trained without the noisy training to demonstrate the deterioration caused by the distribution shift. Following [64], we dropped

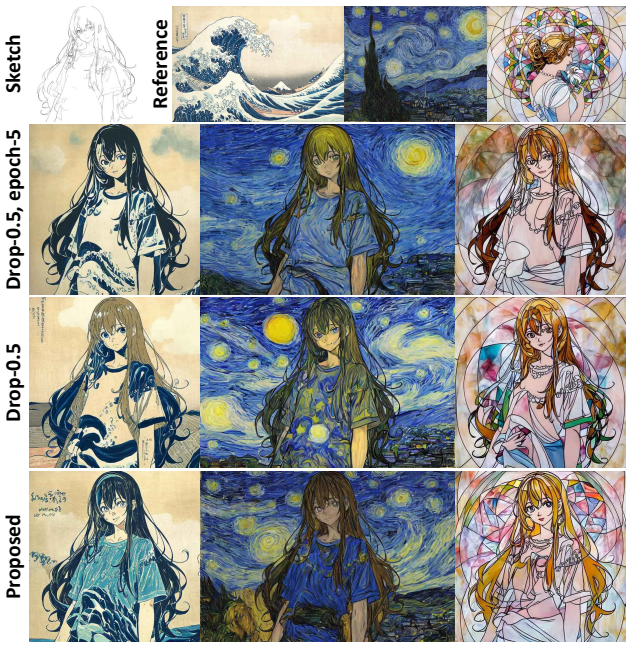

Figure 7. Comparison with ablation models to demonstrate the influence of training duration on style transfer.

50% prompt inputs during training, which are reference images in our task. This model is labeled as *Drop-0.5*.

An alternative solution to reduce spatial entanglement is adopting the CLS token as prompt input instead of local tokens. As the CLS token is globally compressed, it contains much less spatial information. The following ablation models utilize the CLS token in two distinct ways:

2. Projection model: CLS token is decomposed into 256 heads through two linear layers with an in-between activation. This decomposition occurs before the token is input into the denoising U-Net. This model is labeled as *Proj-0.1*.

3. CLS model: Since the CLS token is a 256-dimension vector, we replace cross-attention modules with linear layers to reduce computational cost. This model is labeled as *CLS-0.1*.

Discussion. A quantitative evaluation measured by 50K-FID is shown in Table 1. Notably, the inferior scores of *Proj-0.1* and *CLS-0.1* models suggest that the CLS token is less effective than local tokens for training reference-based models. Besides, the spatial entanglement still appears in these models as the CLS token also contains enough spatial information to reconstruct images, inferable from IP-Adapter [57]. Therefore, we consider local tokens a better choice as reference embeddings.

For the *Drop-0.5* model, we calculated its FIDs at two different epochs to demonstrate the deterioration in image quality caused by the distribution shift, which intensifies as training progresses, as discussed in Section 3.2.

The FID results of *Drop-0.5* model at the 5th epoch are closer to that of the proposed model and better than those calculated at 7th epoch. Therefore, we compare the model at 5th epoch for spatial entanglement, as illustrated in Figure 6, where the results of the *Drop-0.5* model are still inferior to the proposed model trained for seven epochs. As is demonstrated in Figure 7, the results of two models trained for seven-epoch have more fine-grained textures and stories, indicating that longer training is necessary for improving style transfer performance.

Table 2. FID comparison between the proposed model and major baseline methods. We utilized Karras noise scheduler in this test [23]. Notably, the comparison of T2I results suggests that text-based generation is also affected by the distribution shift. "CN": ControlNet; †: Texts were paired with mismatched sketch images to examine the distribution shift in the T2I model.

50K-FID ↓	
Reference-based	
Ours, GS-5	**5.5272**
CN-Lineart + SD v1.5 + IP-Adapter-vitH	25.8390
CN-Lineart + SD v1.5 + IP-Adapter-vitG	27.7849
CN-Anime + Anything v3 + IP-Adapter-ft	23.2523
CN-Anime + Anything v3 + IP-Adapter-vitH	39.2049
CN-Anime + Anything v3 + IP-Adapter-vitG	27.5994
CN-Anime + Anything v3 + Self-injection	21.0125
AnimeDiffusion [3]	62.3451
Yan et al. [56]	26.1816
Text-based	
CN-Anime + Anything v3	20.1411
†CN-Anime + Anything v3	27.4624

4.3. Comparison with baseline

Existing reference-based methods are developed based on GANs [30, 56] or only for anime faces [3]. Therefore, the effectiveness of these methods is limited by the generative capacity of the network or the failure to address spatial mismatch.

Our major baselines are combinations of ControlNet and IP-Adapter [14, 28, 34, 53, 57, 60, 64, 68] since they are publicly available and have demonstrated effectiveness in generating high-quality images for general purposes. However, as ControlNet and IP-Adapter are designed for T2I generation and I2I re-imagination, respectively, they were not trained with concern for the potential conflicts within spatial semantics. Therefore, these combinations severely suffer from the distribution shift when applied to our task.

We adopted two variations of LDMs in this evaluation: *SD v1.5* [42, 45] and *Anything v3* [58]. We focus on *Anything v3* as it is personalized for anime-style images and serves as the SD backbone for training *ControlNet-Linaert-anime*, according to the official document of [62]. We omit SDXL and its variations as we found related combinations ineffective in generating satisfying results. Quality-related

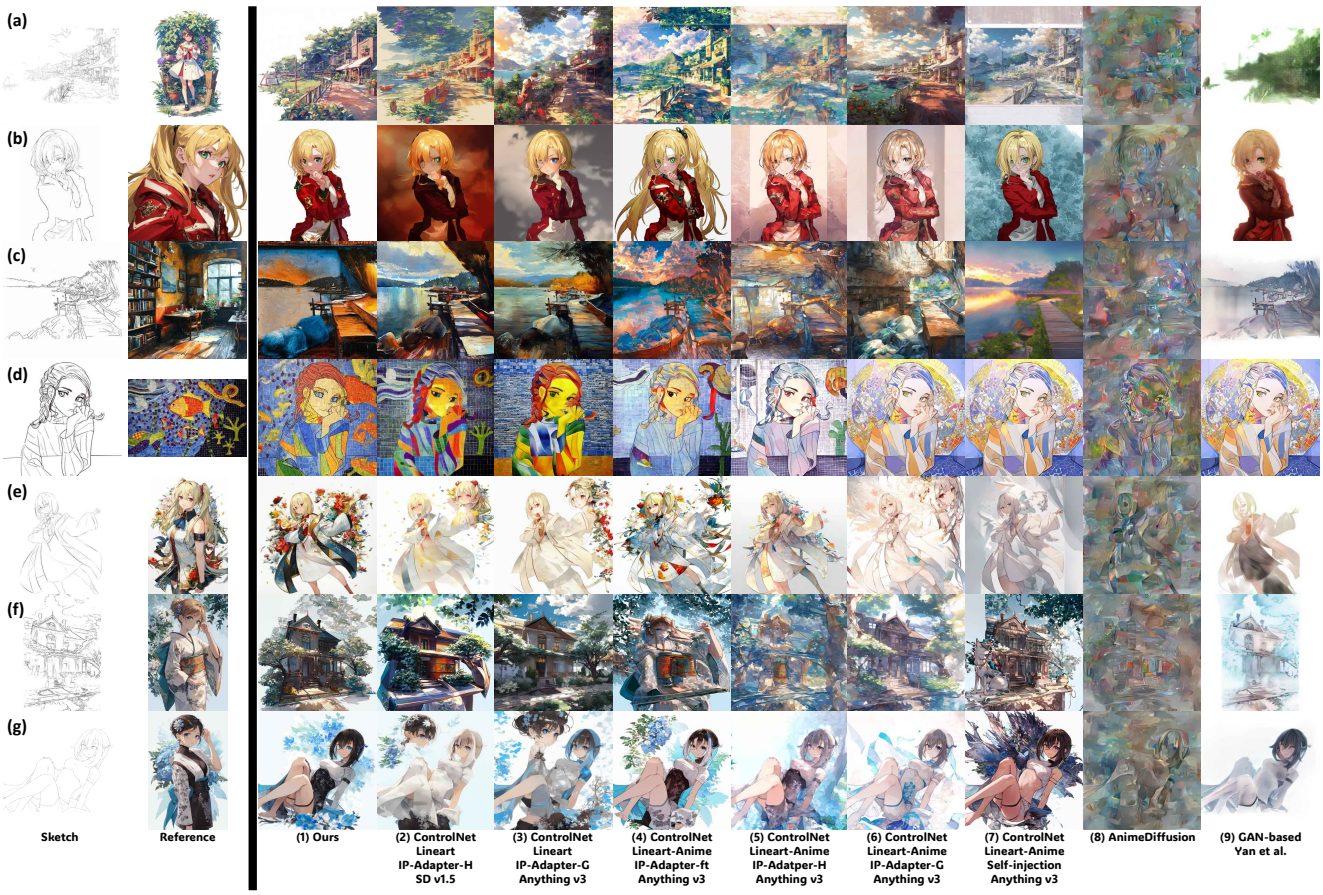

Figure 8. Qualitative comparison with baseline methods. Adapter-based baseline results were generated using different CASs that stress the transfer of style rather than composition, while ours were fixed as 1. Combinations are labeled as {Control condition adapter}-{Control condition}-{Image prompt adapter}-{SD model} from top to bottom. Zoom in for details.

Figure 9. Examples of artifacts selected from Figure 8.

prompts were adopted in all testing, such as "masterpiece, best quality" and so on [16, 44].

Note that we fine-tuned the *IP-Adapter v1.5* with *Anything v3* on our training set for five epochs to ensure both adapters are well-trained for anime-style images, with the fine-tuned adapter labeled as *IP-Adatper-ft*. Additional qualitative comparisons with more baselines are included in the supplementary materials.

Quantitative comparison. Table 2 lists the FID scores of major baselines and demonstrates the superiority of our model in reference-based sketch colorization. We attribute this advancement to the improvement in style transfer. Noticeably, we calculated the FIDs of the T2I sketch colorization to highlight the considerable impact of distribution shift on perceptual quality. The inferior result was achieved using semantically misaligned texts as prompts, which simulate the common cases of reference-based colorization.

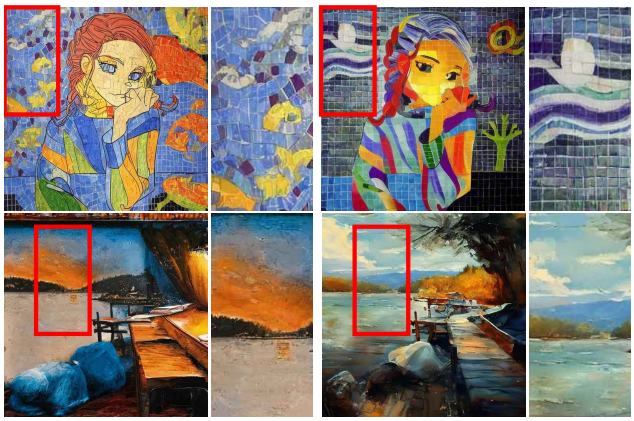

Figure 10. Examples selected from Figure 8 to highlight the quality of transferred textures and styles.

Qualitative comparison. A qualitative comparison is given in Figure 8, where [56] failed to generate acceptable

images with such inputs. Results of adapter combinations were generated using a set of CASs suggested by [54] to emphasize style transfer and downplay composition/content transfer. This corresponds to the 'strong style transfer' setting in WebUI [2].

We select several results from Figure 8 to highlight the conflicting parts, as shown in Figure 9, where 1) The character was incorrectly generated in the house sketches, 2) Long hair was generated in the short-hair sketches, and 3) Redundant hair and clothes appeared outside the character. We also emphasize the improvement in texture generation in Figure 10, where the proposed model more effectively transferred fine-art textures in the backgrounds than the baseline method.

User study. Given that existing metrics cannot estimate the distribution shift or the semantic similarity between generated results and reference images in sketch colorization, we conducted a user study for subjective evaluation.

We selected three baseline methods that achieved top results in the FID evaluation and invited 20 participants to assess each method across six dimensions after testing them, and most participants are familiar with DMs. The six dimensions include (a) Overall rating; (b) Perceptual quality of generated results, estimating whether the generated images are visually pleasant; (c) Correctness of transfer, determining whether incompatible semantics are filtered out; (d) Style similarity with references, evaluating similarity regarding color, texture, and stroke; (e) Semantic fidelity to sketches, checking whether segmentation of results follows that of sketch inputs; and (f) Easiness of achieving satisfying results, noting adjustments and re-generations before achieving a satisfying result. The initial settings of hyperparameters for baseline methods were the same as the qualitative evaluation.

We prepared instructions and a video to clarify the questions and how to generate optimal results for each method. Participants were required to test at least ten groups of inputs, covering four types of pairs: 1) figure sketch with figure reference, 2) figure sketch with non-figure reference, 3) non-figure sketch with figure reference, and 4) non-figure sketch with non-figure reference. We set the batch size to 4 during the testing and allowed up to 10 re-generations for each input pair, ensuring participants checked over forty results from each method before rating. The participants could choose any images from our test dataset or their own data, but all reference images had to be fine art images. Responses were collected anonymously.

Results of the user study are visualized in Figure 11, where higher scores across all six dimensions indicate that the proposed model is preferable to the baseline methods, owing to a significant improvement in image quality and similarity, as well as a much lower probability of spatial entanglement. More detailed visualizations are included in the supplementary materials.

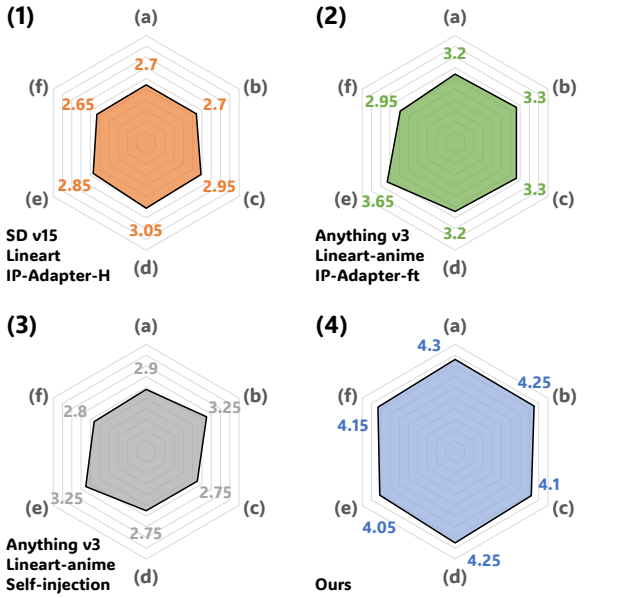

Figure 11. Visualization of user study results. Users were required to rate each method from six dimensions: (a) overall performance, (b) perceptual quality of results, (c) correctness of transfer, (d) style similarity with references, (e) semantic fidelity to sketches, (f) easiness of achieving satisfying results. Higher scores indicate better performance.

5. Conclusion

In this paper, we comprehensively investigated the distribution shift, a critical issue in reference-based sketch colorization, and proposed a two-stage training strategy with a novel training method, termed "noisy training," to diminish the impact of this problem. Our qualitative/quantitative evaluations and the user study validated the superiority of the proposed model in image quality, similarity, and semantic fidelity to sketches.

Nevertheless, there are still some limitations remain to be further discussed for this work: 1. the segmentation of results deteriorates when combined with attention injection; 2. the distribution shift is mitigated but not fully solved, spatial entanglements still appear in corner cases of generated results by our proposed model.

6. Acknowledgement

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (A) No. 21H04916. We extend our gratitude to Xinrui Wang for his invaluable assistance with the writing and to Fuminori Shibasaki and Ayumu Sato for their contributions of hand-drawn sketches.

References

[1] Kenta Akita, Yuki Morimoto, and Reiji Tsuruno. Colorization of line drawings with empty pupils. *Comput. Graph. Forum*, 39(7):601–610, 2020. 2

[2] Automatic1111. stable-diffusion-webui. https://github.com/AUTOMATIC1111/stable-diffusion-webui/tree/master, 2023. Accessed: DATE 2023-06-25. 8

[3] Yu Cao, Xiangqiao Meng, P. Y. Mok, Tong-Yee Lee, Xueting Liu, and Ping Li. Animediffusion: Anime diffusion colorization. *TVCG*, pages 1–14, 2024. 2, 3, 5, 6

[4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 2

[5] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023. 3

[6] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797. IEEE/CVF, 2018. 2

[7] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, pages 8185–8194. IEEE/CVF, 2020. 2

[8] Danbooru community, Gwern Branwen, and Anonymous. Danbooru2021: A large-scale crowdsourced and tagged anime illustration dataset. https://gwern.net/danbooru2021, 2022. Accessed: DATE 2022-01-21. 5

[9] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 3

[10] 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. *CoRR*, 2024. 4

[11] Sébastien Fourey, David Tschumperlé, and David Revoy. A fast and efficient semi-guided algorithm for flat coloring line-arts. In *Vision, Modeling and Visualization VMV*, pages 1–9. Eurographics Association, 2018. 3

[12] Chie Furusawa, Kazuyuki Hiroshiba, Keisuke Ogaki, and Yuri Odagiri. Comicolorization: semi-automatic manga colorization. In *SIGGRAPH Asia*, pages 12:1–12:4. ACM, 2017. 3

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014. 2

[14] h94. Hugging face/ip-adapter. https://huggingface.co/h94/IP-Adapter, 2024. Accessed: DATE 2024-01-02. 6

[15] Reimu Hakurei. Hugging face/waifu-diffusion-v1-4. https://huggingface.co/hakurei/waifu-diffusion-v1-4, 2023. Accessed: DATE 2023-03-05. 3

[16] Havoc. Easynegative. https://civitai.com/models/7808/easynegative, 2023. Accessed: DATE 2023-02-10. 7

[17] Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. Deep exemplar-based colorization. *ACM Trans. Graph.*, 37(4):47, 2018. 3

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pages 6626–6637, 2017. 5

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. 3, 4

[21] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. https://github.com/mlfoundations/open_clip, jul 2021. 3

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976. IEEE/CVF, 2017. 2, 3

[23] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *NeurIPS*, 2022. 3, 6

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410. IEEE/CVF, 2019. 2

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8107–8116. IEEE/CVF, 2020. 2

[26] Hyunsu Kim, Ho Young Jhoo, Eunhyeok Park, and Sungjoo Yoo. Tag2pix: Line art colorization using text tag with secat and changing loss. In *ICCV*, pages 9055–9064. IEEE/CVF, 2019. 3

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[28] Kohya-ss. Hugging face/controlnet-lllite. https://huggingface.co/kohya-ss/controlnet-lllite, 2024. Accessed: DATE 2024-01-02. 3, 6

[29] Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In *CVPR*, pages 5800–5809. IEEE/CVF, 2020. 5

[30] Zekun Li, Zhengyang Geng, Zhao Kang, Wenyu Chen, and Yibo Yang. Eliminating gradient conflict in reference-based line-art colorization. In *ECCV*, pages 579–596. Springer, 2022. 2, 3, 5, 6

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*. OpenReview.net, 2019. 5

[32] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022. 3

[33] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *CoRR*, abs/2211.01095, 2022. 3

[34] Lyumin Zhang Mikubill. sd-webui-controlnet. https://github.com/Mikubill/sd-webui-controlnet, 2023. Accessed: DATE 2023-07-01. 3, 6

[35] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *CoRR*, abs/2302.08453, 2023. 2, 3, 4

[36] Amal Dev Parakkat, Pooran Memari, and Marie-Paule Cani. Delaunay painting: Perceptual image colouring from raster contours with gaps. *Computer Graphics Forum*, 41(6):166–181, 2022. 3

[37] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4172–4182. IEEE, 2023. 2

[38] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952, 2023. 2, 3

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pages 8748–8763. PMLR, 2021. 3

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. 2, 3

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE/CVF, 2022. 2, 3, 6

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, volume 9351, pages 234–241. Springer, 2015. 2

[44] rqdwdw. negativexl. https://civitai.com/models/118418/negativexl, 2023. Accessed: DATE 2023-02-10. 7

[45] runwayml. stable-diffusion-v1-5. https://huggingface.co/runwayml/stable-diffusion-v1-5, 2024. Accessed: DATE 2024-01-02. 6

[46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 3

[47] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, 2023. Accessed: DATE 2023-05-17. 5

[48] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, volume 37, pages 2256–2265. JMLR.org, 2015. 2

[49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*. OpenReview.net, 2021. 3

[50] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*. OpenReview.net, 2021. 2, 3, 5

[51] Daniel Sýkora, John Dingliana, and Steven Collins. Lazybrush: Flexible painting tool for hand-drawn cartoons. *Comput. Graph. Forum*, 28(2):599–608, 2009. 3

[52] TencentARC. Hugging face/ip-adapter. https://github.com/TencentARC/T2I-Adapter/tree/SD, 2024. Accessed: DATE 2024-01-02. 3

[53] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, pages 1921–1930. IEEE/CVF, 2023. 6

[54] Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 8

[55] Xiaoyu Xiang, Ding Liu, Xiao Yang, Yiheng Zhu, Xiaohui Shen, and Jan P. Allebach. Adversarial open domain adaptation for sketch-to-photo synthesis. In *WACV*, pages 944–954. IEEE/CVF, 2022. 5

[56] Dingkun Yan, Ryogo Ito, Ryo Moriai, and Suguru Saito. Two-step training: Adjustable sketch colourization via reference image and text tag. *Computer Graphics Forum*, 2023. 2, 3, 5, 6, 7

[57] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *CoRR*, abs/2308.06721, 2023. 2, 3, 6

[58] Yuno779. https://civitai.com/models/9409, 2023. Accessed: DATE 2023-06-25. 6

[59] Lvmin Zhang. Sketchkeras. https://github.com/lllyasviel/sketchKeras, 2017. 5

[60] Lvmin Zhang. How controlnet-reference works. https://github.com/Mikubill/sd-webui-controlnet/discussions/1236, 2023. 6

[61] Lvmin Zhang. Style2paints v5, 2023. Accessed: DATE 2023-06-25. 3

[62] Lvmin Zhang. Controlnet-v1-1-nightly. https://github.com/lllyasviel/ControlNet-v1-1-nightly, 2024. Accessed: DATE 2024-01-02. 3, 6

[63] Lvmin Zhang, Chengze Li, Tien-Tsin Wong, Yi Ji, and Chunping Liu. Two-stage sketch colorization. *ACM Trans. Graph.*, 37(6):261, 2018. 2, 3

[64] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2, 3, 5, 6

[65] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *ECCV*, volume 9907, pages 649–666. Springer, 2016. 3

[66] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, and Alexei A. Efros. Real-time user-guided image colorization with learned deep priors. *ACM Trans. Graph.*, 36(4):119:1–119:11, 2017. 3

[67] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Trans. Graph.*, 42(6):244:1–244:14, 2023. 4

[68] Yuechen Zhang, Jinbo Xing, Eric Lo, and Jiaya Jia. Real-world image variation by aligning diffusion inversion chain. 2023. 6

[69] Changqing Zou, Haoran Mo, Chengying Gao, Ruofei Du, and Hongbo Fu. Language-based colorization of scene sketches. *ACM Trans. Graph.*, 38(6), 2019. 3