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

ANONYMOUS AUTHOR(S)
SUBMISSION ID: 1021

Fig. 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. References (3)-(5): ©tarotaro.

Diffusion models have recently demonstrated their effectiveness in generating extremely high-quality images and are utilized in a wide range of applications, including automatic sketch colorization. Although many methods have been developed for this purpose, few have explored the potential conflicts between image prompts and sketch inputs, which often cause severe deterioration in the results. Therefore, this paper investigates latent diffusion models that aim to colorize sketch images using reference color images, especially for a major shortcoming compared to text-based counterparts, termed "distribution problem." We exhaustively analyze this problem and propose a new training strategy to diminish its influence. Comprehensive evaluations with ablation models and baseline methods are conducted to demonstrate the superiority of the proposed methods in reference-based sketch colorization.

XXXX-XXXX/2024/5-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS Concepts: • **Applied computing** → **Fine arts**; • **Computing methodologies** → **Computer vision**; *Image processing*.

Additional Key Words and Phrases: Sketch colorization, Dual-conditioned generation, Latent diffusion model, Latent manipulation

ACM Reference Format:
Anonymous Author(s). 2024. ColorizeDiffusion: Improving Reference-based Sketch Colorization with Latent Diffusion Model. 1, 1 (May 2024), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Anime-style images have gained worldwide popularity over the past few decades thanks to their diverse color composition and captivating character design, but the process of colorizing sketch images has remained labor-intensive and time-consuming. However, swift advancements in diffusion models [Ho et al. 2020; Zhang et al. 2023b] now enable large generative models to create remarkably high-quality images across various domains, including anime style. Most conditional diffusion models predominantly focus on Text-to-Image (T2I) generation, but few specialize in the reason for the deterioration when applying image-guided models to reference-based sketch colorization, a complex dual-conditioned generation

task that utilizes both a reference and a sketch image. As such, this paper focuses on reference-based colorization and thoroughly analyzes the reasons for this deterioration.

Text-based models, despite their advantages, show a major limitation in transferring attributes from reference images [Hu et al. 2022; Rombach et al. 2022; Ruiz et al. 2023]. Therefore, researchers have proposed many methods to introduce image prompts [Mou et al. 2023; Ye et al. 2023] to pre-trained T2I models through separately trained adapters. Though these adapters are effective in reimagining or reconstructing, they suffer from a deterioration in perceptual quality due to the misalignment of the distributions when combined with other adapters and cannot handle conflicts between image conditions effectively, which is a critical issue in reference-based sketch colorization.

Both sketch and reference images contain varied information about structure, layout, and identity, potentially expressing incompatible contents. This issue, termed the "distribution problem" in this paper, stems from the semantic alignment of training data, where reference images consistently align semantically with the ground truth, leading networks to prioritize reference embeddings over sketch semantics during inference. We propose a two-stage training to diminish this issue. Additionally, we introduce specific strategies to further mitigate this problem. The investigation and solution to the distribution problem constitute the key points of this paper.

Through rigorous experimentation with ablation models and baselines, we empirically proved the effectiveness of the proposed methods in reference-based colorization. We also conducted a subjective user study to evaluate the proposed method.

The contribution of this paper lies in the analysis of condition conflicts in reference-based sketch colorization and the two-stage training strategy proposed to diminish its influence. This strategy includes a novel training method, namely noisy training, designed to improve the quality of results. We also developed a zero-shot text-based manipulation method and a user interface to strengthen the controllability and interactivity of the proposed model, with further details provided in the supplementary materials. We will release our code and pre-trained models.

2 RELATED WORK

Our work focuses on reference-based sketch colorization, an important subfield of image generation. We utilize the score-based generative model [Ho et al. 2020; Rombach et al. 2022; Song et al. 2021b] as our neural backbone, which is widely known as the diffusion model. Our training methods and overall pipeline are designed following previous style transfer [Gatys et al. 2016; Isola et al. 2017] and colorization methods, pursuing pixel-level correspondence and fidelity to the input sketch image.

Latent Diffusion Models. Diffusion probabilistic Models (DMs) [Ho et al. 2020] are a class of latent variable models inspired by considerations from nonequilibrium thermodynamics [Sohl-Dickstein et al. 2015]. Compared with Generative Adversarial Nets (GANs) [Choi et al. 2018, 2020; Goodfellow et al. 2014; Karras et al. 2019, 2020], DMs excel at generating highly realistic images across various contexts. However, the autoregressive denoising process, typically computed using a U-Net network [Ronneberger et al. 2015], incurs substantial computational costs. To address this limitation, Stable Diffusion [Podell et al. 2023; Rombach et al. 2022], a typical class of Latent Diffusion Models (LDMs), utilizes a two-stage synthesis and carries out the diffusion/denoising process within a highly compressed latent space to reduce computational costs significantly. Concurrently, many efficient samplers have been proposed to accelerate the denoising process [Lu et al. 2022a,b; Song et al. 2021a,b]. In this paper, we adopt a pre-trained text-based SD model as our neural backbone, utilize DPM++ solver and Karras noise scheduler [Karras et al. 2022; Lu et al. 2022b; Song et al. 2021b] as the default sampler, and employ classifier-free guidance [Dhariwal and Nichol 2021; Ho and Salimans 2022] to strengthen the reference-based performance.

Neural Style Transfer. First proposed in [Gatys et al. 2016], Neural Style Transfer (NST) has now become a widely adopted technique compatible with many effective generative models. Reference-based colorization, which aims to transfer semantics, color, and textures from reference images to sketch images, can be viewed as a subclass of multi-domain style transfer. However, compared to other NST methods [Choi et al. 2018, 2020; Huang and Belongie 2017; Johnson et al. 2016; Zhang et al. 2023a; Zhu et al. 2017], colorization requires a higher level of correspondence in both color and textures with the reference while maintaining semantic fidelity to the sketch inputs. Consequently, our method is developed based on the principles of conditional image-to-image translation [Isola et al. 2017] to ensure pixel-level correspondence between the sketch and colorized results.

Sketch Colorization. Many effective methods have been developed to achieve automatic sketch colorization, all of which can be divided into traditional [Fourey et al. 2018; Furusawa et al. 2017; Parakkat et al. 2022; Sýkora et al. 2009] or Deep Learning (DL)-based methods [He et al. 2018; Isola et al. 2017; Zhang et al. 2016]. Owing to the swift advancement in generative networks, DL-based methods can generate much better results than traditional methods.

According to guiding conditions, existing DL-based methods can be categorized into three types: text-based [Kim et al. 2019; Zhang et al. 2023b; Zou et al. 2019], user-guided [Zhang et al. 2018, 2017], and reference-based [Akita et al. 2020; Lee et al. 2020; Sun et al. 2019; Yan et al. 2023]. Text-based methods are the most popular subclass nowadays, owing to sufficient community-developed T2I DMs, as well as many plug-in adapters and mini-scale fine-tuning methods [Hu et al. 2022; Ruiz et al. 2023; Ye et al. 2023; Zhang et al. 2023b]. However, text-based models cannot effectively handle the conflicts between image conditions as plug-in modules are trained with respective conditions. User-guided methods require users to have a basic knowledge of line art [Zhang et al. 2018] and become inefficient for batch processing. Compared to other reference-based tasks, such as recoloring grayscale images, sketch colorization is more challenging because sketch images lack semantics on fine-grained textures and deterministic information on identities, making the networks easily transfer incompatible semantics after training. Limited by such conflicts and the generative ability of networks, existing reference-based methods are only developed for small-scale tasks, such as figure-to-figure or face-to-face tasks [Cao et al. 2024; Choi et al. 2020; Lee et al. 2020; Li et al. 2022; Yan et al. 2023]. To

Fig. 2. Illustration of semantic conflicts in T2I colorization. The network prioritizes prompt conditions over the sketch in the arm regions. This preference results in unexpected colorization discrepancies, particularly in areas anticipated to be skin-toned, thereby leading to visually discordant segmentation. Presented results are derived from the *ControlNet_lineart_anime + Anything v3* framework. ©Style2paints.

Fig. 3. Illustration of the spatial entanglement and also a comparison of diminishing the distribution problem manually by users and automatically by the proposed model. ©SDAI.

train a generalized model, we investigate the condition conflict and propose a novel two-stage training method.

3 REFERENCE-BASED COLORIZATION

This section outlines the reference-based training, the "distribution problem" encountered in dual-conditioned generation with DMs, and our proposed solutions. We use r and s for reference and sketch inputs, respectively.

Reference-based sketch colorization aims to transfer style and compatible content semantics from references to sketches without changing the original semantics of sketches. However, it is challenging for DMs to adaptively distinguish between compatible and incompatible content in image-guided models, because most content semantics are jointly transferred with style semantics at the same level using the same module, leading to a deterioration we defined as the "distribution problem" in this paper.

We use ϵ for random noise, z_t for latent representations at timestep t, and y for ground truth. The pre-trained encoder, decoder, and the denoising U-Net are denoted as $\mathcal{E}$, $\mathcal{D}$, and θ, respectively. The timestep t starts from $T-1$ to 0, where T is the total diffusion step set to 1000.

3.1 Distribution Problem

We introduce a challenge in automatic sketch colorization using DMs, which significantly degrades the perceptual quality of generated results and is identified as the "distribution problem" in this paper. We illustrate a T2I example in Figure 2, as texts are more straightforward than image prompts. Unlike text- or user-guided colorization, image-guided methods often involve such conflicts in the reference, leading to severe deterioration in visual outcomes by synthesizing numerous incompatible semantics. We explain this problem from two different aspects to facilitate understanding.

1. DMs "degrade" into a decoder of the reference encoder. Since sketch images have many blank regions, it is challenging for DMs to precisely reconstruct the ground truth based on sketches, especially for complicated backgrounds and compositions, as sketches lack deterministic embeddings compared to grayscale images or canny images [Kohya-ss 2024; Zhang et al. 2023b]. Meanwhile, image embeddings contain much more information than text embeddings, hindering the optimization of networks from learning to synthesize and compose elements naturally since detailed compositions and fine-grained textures are already provided by image embeddings. Training the entire network with both conditions significantly diminishes conflicts and improves image quality, which our comparison with baselines will demonstrate.

2. The underlying reason stems from the distribution level. Given $p(z|y)$, the ground truth distribution, and $p(z|s)$ and $p(z|r)$, two ideal conditional distributions, we assume images composed of features only from $p(z|s)$ (or only from $p(z|r)$) are visually pleasant color images. Therefore, if all features of colorized images are sampled from $p(z|s)$, the quality, segmentation, and fidelity to sketches of results will not be degraded by references r. However, since $p(z|r)$ semantically aligns with $p(z|y)$ during training, the networks become ineffective in disentangling sketch semantics from references, and the optimized $p_\theta(z|s,r)$ always deviates from $p(z|s)$. Consequently, $p_\theta(z|s,r)$ easily contains contents conflicting with sketches, making the model unable to generate visually pleasant texture and content. Moreover, image prompts implicitly express size- and layout-related embeddings [Podell et al. 2023], which are likely to degrade the perceptual quality of reference-based results if accurately transferred to incompatible sketches. Our ablation study will demonstrate this deterioration.

For better visualization and comparison, we define a subclass of semantic conflict as "spatial entanglement," which erroneously generates incompatible content outside sketches, such as changing the hairstyle or synthesizing redundant identities/objects. Spatial entanglement widely exists in results generated using portraits as references, and an example is illustrated in Figure 3.

A feasible trade-off is manually adjusting sampling hyperparameters during inference to enhance sketch fidelity. However, such adjustments are less effective for spatial entanglement and generally degrade the style similarity with reference images. Furthermore, finding optimal combinations of hyperparameters with existing methods is extremely difficult as they are combinations of independently trained adapters.

3.2 Architecutre

The proposed model comprises a pre-trained VAE $\mathcal{E}$, $\mathcal{D}$, a sketch encoder, a denoising U-Net θ, and a pre-trained CLIP Vision Transformer (ViT) from OpenCLIP-H [Cherti et al. 2023; Ilharco et al. 2021; Radford et al. 2021; Schuhmann et al. 2022]. The ViT remains frozen during training and outputs 256 local tokens, which serve as reference embeddings for generation. The architecture of the denoising U-Net is similar to SD v2.1 [Rombach et al. 2022]. Notably, the model is initialized using Waifu Diffusion [Hakurei 2023], except for the cross-attention modules, which are trained from scratch.

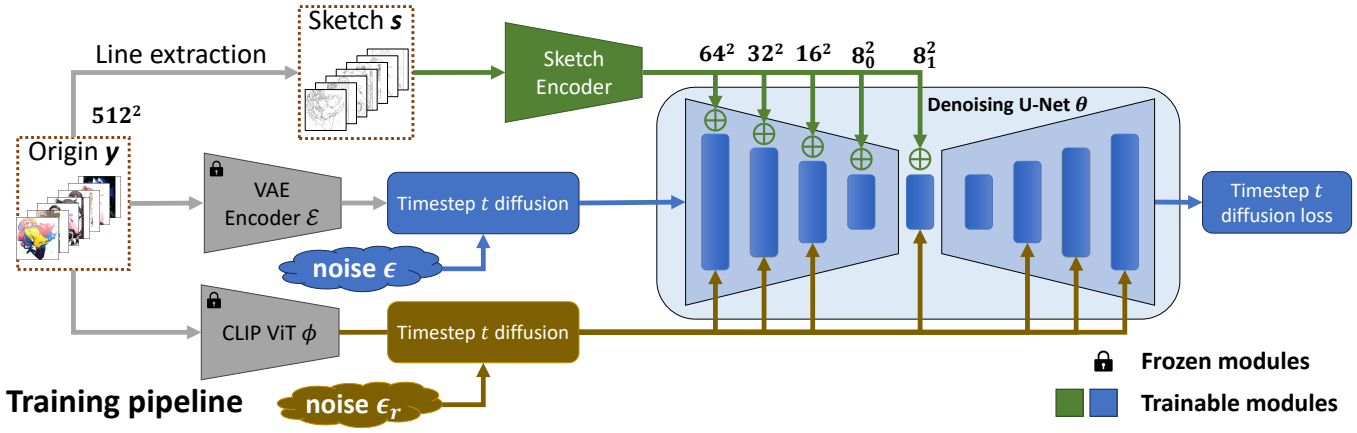

Fig. 4. The architecture of our colorization model and the pipeline of the proposed noisy training. The noisy training is designed to optimize the style transfer and hinder the optimization of content transfer by adding timestep-dependent noise to image embeddings during training.

We utilize a multi-layer sketch encoder to enhance the controllability of sketches in semantics. Extracted sketch features are input into the encoder of the denoising U-Net at all scales by being added to the forward features, as illustrated in Figure 4. We can adjust the scale of added sketch features. Given $size \in \{64^2, 32^2, 16^2, 8_0^2, 8_1^2\}$, this addition is formulated as $\hat{z}^{size} = z^{size} + \lambda_s s^{size}$, where z^{size} and s^{size} denote forward features and extracted sketch features, respectively. Here, λ_s is defined as the sketch control scale, a hyperparameter that adjusts the fidelity of results to sketch inputs.

We denote the CLIP ViT and extracted tokens as ϕ and τ_ϕ. Our model is trained for 512^2 images and diffuses perceptually-compressed inputs at a size of 64^2.

3.3 Solutions to diminish the distribution problem

To fully solve the distribution problem, the optimized network should automatically filter incompatible semantics from image prompts when given sketches. However, we found it challenging in practice, as score-based models are designed to express all conditions.

Instead, we mitigate the negative impact of the distribution problem through specific training tricks. We propose a two-stage training strategy to hinder the optimization of content transfer, extending the training to improve the style transfer performance without causing severe deterioration in perceptual quality.

Noisy training. As previously analyzed in Section 3.1, longer training leads to a higher probability of generating incompatible content. Many studies [Esser et al. 2024; Mou et al. 2023; Zhang et al. 2023a] have demonstrated that the denoising of DMs is non-uniform. The early sampling steps determine the layout and content by rendering low-level features such as color spots, while subsequent steps refine these spots into detailed objects, identities, textures, and strokes.

Inspired by these findings, we mitigate this issue by diffusing reference embeddings with timestep-dependent noise during training. Given r the reference input and α_t, β_t the hyperparameters of the diffusion 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. Compared to non-uniform timestep sampling, the noisy training disables the transfer in the early steps, rather than slowing the optimization. The pipeline of noisy training is illustrated in Figure 4. We trained the network for five epochs and dropped 10% of reference inputs to ensure the network is aware of the distribution $p(z|s)$ for classifier-free guidance.

Fig. 5. A comparison of inpainting. The upper result is generated by an ablation model trained without center cropping. ©tarotaro(sketch), ©SDAI(reference)

Second-stage training. The noisy-trained models cannot accurately generate white backgrounds or fine-grained strokes. Therefore, we perform a second-stage fine-tuning to recover the early-step transfer slightly. This fine-tuning follows the vanilla diffusion training. Besides, we drop 50% of reference inputs to avoid the distribution problem and 10% of sketch inputs for classifier-free guidance. Unavoidably, the second-stage training results in severe spatial entanglement if it lasts too long. We empirically set this fine-tuning to two epochs.

Table 1. Quantitative comparison of FIDs with ablation models. We used the variance preserving (VP) noise scheduler [Song et al. 2021b] in this 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, Projection-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
Proposed model	7.3676	**6.8551**

Center cropping. Image-guided networks trained using both conditions show an inability in inpainting. We assume it is caused by their sensitivity to sketch inputs and view-related embeddings, which are implicitly expressed by image prompts. An example is shown in Figure 5, where a girl and doors are generated in front of the house, making the result less satisfying, though semantically correct. Consequently, we applied center cropping to sketch inputs during training so the network learned to generate perceptually pleasant content in the margins. However, due to the existence of view-related embeddings, the effectiveness of center cropping diminishes without the proposed noisy training. This is the second reason for restricting the duration of second-stage fine-tuning.

Specifically, image deformation has been widely used to produce reference images as training data in previous reference-based sketch colorization methods [Cao et al. 2024; Lee et al. 2020; Yan et al. 2023]. Yet, we found that it degrades the quality of the generated results without providing a notable improvement in avoiding spatial entanglement for DMs in pre-experiments. Therefore, we discard this augmentation.

4 EXPERIMENT

In this section, we compare the proposed model with the ablation models in Section 5.1 and the baseline methods in Section 5.2. For simplicity, we denote the reference-based classifier-free guidance scale [Ho and Salimans 2022] as 'GS' and the cross-attention scale as 'CAS'. Implementation details are included in the supplementary materials.

To quantitatively estimate the perceptual quality of generated images, we utilize the Fréchet Inception Distance (FID) [Heusel et al. 2017; Seitzer 2023], which quantifies the distance between the distributions of generated results and ground truth.

4.1 Ablation Study and Discussion

Architecture. We introduce three important ablation models here to investigate the distribution problem, 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 a degradation in composition and texture quality. All ablation models were trained for seven epochs as the proposed one.

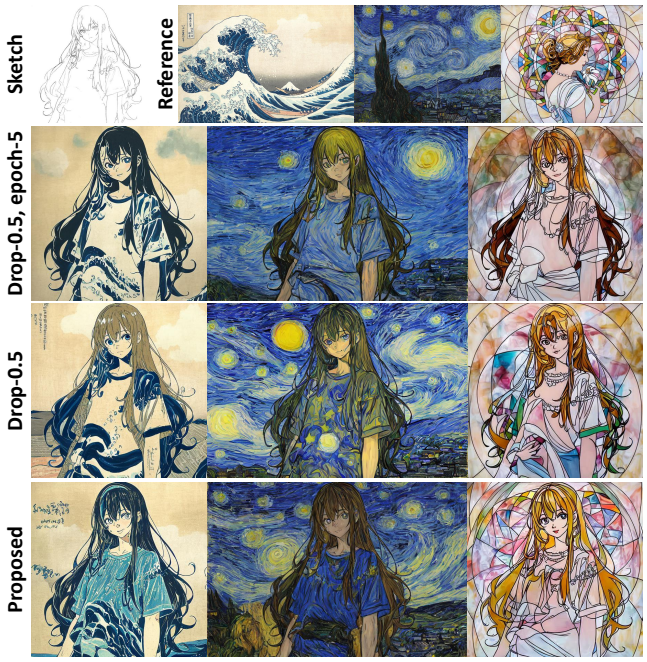

Fig. 6. Comparison with ablation models to demonstrate the influence of training duration on style transfer.

1. Dropping model: We trained this model, labeled as Drop-0.5, without the proposed noisy training to demonstrate and visualize the deterioration caused by the distribution problem. Following [Zhang et al. 2023b], we dropped 50% reference inputs during training. As the content transfer is consistently optimized, it becomes difficult, and sometimes impossible, to maintain style similarity in the generated results to the reference while eliminating spatial entanglement in this model. Other ablation models are labeled in the same format.

An alternative solution to reduce spatial entanglement is adopting the CLS token 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: In this model, the CLS token is decomposed into 256 heads through a trainable projection module, which consists of two linear layers with an in-between activation. This decomposition occurs before the token is input into the denoising U-Net. It is labeled as *Proj-0.1*.

3. CLS model: Since the CLS token is a vector, cross-attention modules are replaced by linear layers to reduce computational cost. This model is labeled as *CLS-0.1*.

Discussion. We first calculated 50K-FID to quantitatively evaluate the perceptual quality, as shown in Table 1. Notably, the inferior scores of *Proj-0.1* and *CLS-0.1* models suggest that the CLS token is much less effective for directly training reference-based models. Besides, the spatial entanglement is not removed in these models as the CLS token also contains enough spatial information to reconstruct images, which is inferable from IP-Adapter [Ye et al. 2023].

Table 2. FID comparison between the proposed model and major baseline methods. We utilized Karras noise scheduler in this test [Karras et al. 2022]. Notably, the comparison of T2I results suggests that text-based generation is also affected by the distribution problem. "CN": ControlNet; †: Texts were paired with mismatched sketch images to examine the distribution problem 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
Text-based	
CN-Anime + Anything v3	20.1411
†CN-Anime + Anything v3	27.4624

Fig. 7. Examples of the distribution problem selected from Figure 11.

Therefore, we chose local tokens as reference embeddings for better texture quality.

For the *Drop-0.5* model, we calculated its FIDs at two different epochs to illustrate the deterioration caused by the distribution problem, which intensifies as training progresses, as discussed in Section 3.1. More specifically, layout, view, and size-related embeddings are more and more accurately transferred in early steps, degrading spatial disentanglement, composition, and texture quality. With the proposed noisy training to mitigate the distribution problem, the two-stage trained model achieved the best score in this evaluation.

The FIDs of the five-epoch *Drop-0.5* model are closer to those of the proposed model and better than those calculated at epoch 7. Therefore, we compare the five-epoch model for spatial entanglement, as illustrated in Figure 10, where the results of the *Drop-0.5* model are still inferior to those of the proposed model, which was trained for seven epochs. We then demonstrate the importance of longer training duration by comparing the style transfer performance of the three models, visualized in Figure 6, where both the seven-epoch *Drop-0.5* model and the proposed model generated more fine-grained textures and strokes.

4.2 Comparison with Baseline

Our major baselines are combinations of ControlNet and IP-Adapter [h94 2024; Kohya-ss 2024; Mikubill 2023; Tumanyan et al. 2023; Ye et al. 2023; Zhang 2023; Zhang et al. 2023b,c] since they are publicly available and have demonstrated efficiency in generating high-quality images for general purposes. Variations of T2I-Adapter [Mou et al. 2023; TencentARC 2024] also meet the requirements, yet

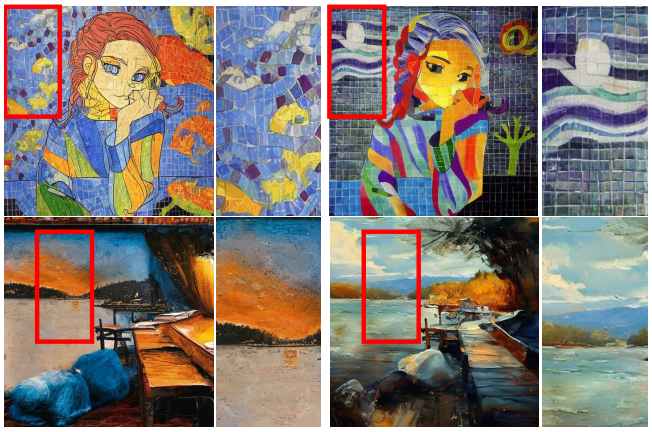

Fig. 8. Examples selected from Figure 11 to highlight the quality of transferred textures and styles.

we found they are less effective in producing reasonable results for sketch colorization. We adopted two variations of LDMs in this evaluation: *SD v1.5* [Rombach et al. 2022; runwayml 2024] and *Anything v3* [Yuno779 2023]. We focus on *Anything v3* as it is personalized for anime-style images and serves as the SD backbone for training the *ControlNet-Linaert-anime* according to the official document of [Zhang 2024]. We omit SDXL and its variations as we found them ineffective in generating satisfying anime-style images. Specifically, we fine-tuned the *IP-Adapter v1.5* with *Anything v3* on our training set for five epochs to align their distributions. The fine-tuned adapter is labeled as *IP-Adapter-ft* in all experiments and included in the supplementary materials with more qualitative comparisons.

Necessary prompts were adopted for models originally designed for T2I generation, such as ("masterpiece, best quality, ultra-detailed, hires") for positive prompts and ("easynegative") [Havoc 2023] or ("negativeXL_D") [rqdwdw 2023] for negative prompts. To avoid the distribution problem, we added "a girl" to the negative prompts when colorizing landscape sketch images with figure images, and to the positive prompts when using landscape images to colorize figure images.

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, as our network is directly trained using image embeddings under both conditions. We also calculated the FIDs for the T2I model, and the significant gap highlights the considerable impact on perceptual quality as the adapter is separately trained.

Qualitative Comparison. We show a qualitative comparison with baseline methods in Figure 11. Baseline results were generated using a set of CASs suggested by [Wang et al. 2024] to emphasize style transfer and downplay composition/content transfer. This corresponds to the 'strong style transfer' setting in WebUI [Automatic1111 2023]. However, for rows (d) and (h), we utilized the

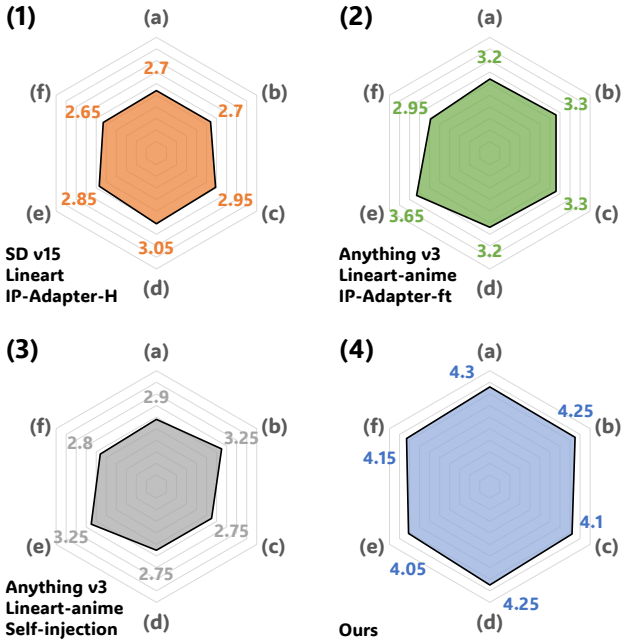

Fig. 9. Visualization of user study results. Users were required to rate each method from six dimensions: (a) overall performance, (b) perceptual quality of results, (c) correctness of transfer, (d) style similarity with references, (e) semantic fidelity to sketches, (f) easiness of achieving satisfying results. A higher score indicates better performance.

'normal' setting as the distribution problem is unlikely with these inputs. To demonstrate the superiority of the proposed model in mitigating the distribution problem, our results were generated using a fixed set of hyperparameters, GS 5 and CASs 1, without any adjustments.

From row (b), we can infer that generating a white background is challenging for image-guided models, as the embedding of "pure white backgrounds" is barely expressed by image prompts and requires a strong transfer ability to express. Specifically, we select several results from Figure 11 to highlight the conflicting parts, as shown in Figure 7, 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 8, where the proposed model more effectively transferred fine-art textures in the backgrounds than the baseline method.

User study. Given that existing metrics cannot effectively estimate the distribution problem or accurately measure the semantic similarity between generated results and reference images in sketch colorization, we conducted a user study to subjectively evaluate our method.

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, the influence of important hyperparameters, and how to generate visually pleasant results for each method. Participants were required to test at least ten pairs of inputs, covering four types of combinations: 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 9, 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.

5 CONCLUSION

In this paper, we comprehensively investigated the distribution problem, a critical issue in reference-based sketch colorization, and proposed a two-stage training strategy that contains 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.

However, our work has three aspects that remain to be further improved: 1. The network is designed to transfer semantics and styles, so it is less effective in reproducing details compared to some reference-based methods; 2. the segmentation of results deteriorates when combined with attention injection; 3. the distribution problem is not fully addressed.

Our future work will extend the proposed model to video colorization and diminish the distribution problem further. We will also try to design a metric for evaluating the distribution problem quantitatively.

REFERENCES

Kenta Akita, Yuki Morimoto, and Reiji Tsuruno. 2020. Colorization of Line Drawings with Empty Pupils. *Comput. Graph. Forum* 39, 7 (2020), 601–610. https://doi.org/10.1111/cgf.14171

Automatic1111. 2023. stable-diffusion-webui. https://github.com/AUTOMATIC1111/stable-diffusion-webui/tree/master. Accessed: DATE 2023-06-25.

Yu Cao, Xiangqiao Meng, P. Y. Mok, Tong-Yee Lee, Xueting Liu, and Ping Li. 2024. AnimeDiffusion: Anime Diffusion Colorization. *TVCG* (2024), 1–14. https://doi.org/10.1109/TVCG.2024.3357568

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

Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *CVPR*. IEEE/CVF, 8789–8797. https://doi.org/10.1109/CVPR.2018.00916

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *CVPR*. IEEE/CVF, 8185–8194. https://doi.org/10.1109/CVPR42600.2020.00821

Prafulla Dhariwal and Alexander Quinn Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*. 8780–8794.

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. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. *CoRR* (2024). https://doi.org/10.48550/ARXIV.2403.03206

Sébastien Fourey, David Tschumperlé, and David Revoy. 2018. A Fast and Efficient Semi-guided Algorithm for Flat Coloring Line-arts. In *Vision, Modeling and Visualization VMV*. Eurographics Association, 1–9. https://doi.org/10.2312/vmv.20181247

Chie Furusawa, Kazuyuki Hiroshiba, Keisuke Ogaki, and Yuri Odagiri. 2017. Comicolorization: semi-automatic manga colorization. In *SIGGRAPH Asia*. ACM, 12:1–12:4. https://doi.org/10.1145/3145749.3149430

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In *CVPR*. IEEE/CVF, 2414–2423. https://doi.org/10.1109/CVPR.2016.265

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *NeurIPS*. 2672–2680.

h94. 2024. Hugging Face/IP-Adapter. https://huggingface.co/h94/IP-Adapter. Accessed: DATE 2024-01-02.

Reimu Hakurei. 2023. Hugging Face/waifu-diffusion-v1-4. https://huggingface.co/hakurei/waifu-diffusion-v1-4. Accessed: DATE 2023-03-05.

Havoc. 2023. EasyNegative. https://civitai.com/models/7808/easynegative. Accessed: DATE 2023-02-10.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPS*. 6626–6637.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*.

Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. *CoRR* abs/2207.12598 (2022). https://doi.org/10.48550/arXiv.2207.12598

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*. OpenReview.net.

Xun Huang and Serge J. Belongie. 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *ICCV*. IEEE/CVF, 1510–1519. https://doi.org/10.1109/ICCV.2017.167

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. 2021. *OpenCLIP*. https://doi.org/10.5281/zenodo.5143773

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *CVPR*. IEEE/CVF, 5967–5976. https://doi.org/10.1109/CVPR.2017.632

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *ECCV*, Vol. 9906. Springer, 694–711. https://doi.org/10.1007/978-3-319-46475-6_43

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the Design Space of Diffusion-Based Generative Models. In *NeurIPS*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.).

Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*. IEEE/CVF, 4401–4410. https://doi.org/10.1109/CVPR.2019.00453

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *CVPR*. IEEE/CVF, 8107–8116. https://doi.org/10.1109/CVPR42600.2020.00813

Hyunsu Kim, Ho Young Jhoo, Eunhyeok Park, and Sungjoo Yoo. 2019. Tag2Pix: Line Art Colorization Using Text Tag With SECat and Changing Loss. In *ICCV*. IEEE/CVF, 9055–9064. https://doi.org/10.1109/ICCV.2019.00915

Kohya-ss. 2024. Hugging Face/controlnet-lllite. https://huggingface.co/kohya-ss/controlnet-lllite. Accessed: DATE 2024-01-02.

Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo. 2020. Reference-Based Sketch Image Colorization Using Augmented-Self Reference and Dense Semantic Correspondence. In *CVPR*. IEEE/CVF, 5800–5809. https://doi.org/10.1109/CVPR42600.2020.00584

Zekun Li, Zhengyang Geng, Zhao Kang, Wenyu Chen, and Yibo Yang. 2022. Eliminating Gradient Conflict in Reference-based Line-Art Colorization. In *ECCV*. Springer, 579–596.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022a. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. In *NeurIPS*.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022b. DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models. *CoRR* abs/2211.01095 (2022). https://doi.org/10.48550/arXiv.2211.01095

Lyumin Zhang Mikubill. 2023. sd-webui-controlnet. https://github.com/Mikubill/sd-webui-controlnet. Accessed: DATE 2023-07-01.

Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. 2023. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. *CoRR* abs/2302.08453 (2023). https://doi.org/10.48550/ARXIV.2302.08453

Amal Dev Parakkat, Pooran Memari, and Marie-Paule Cani. 2022. Delaunay Painting: Perceptual Image Colouring from Raster Contours with Gaps. *Computer Graphics Forum* 41, 6 (2022), 166–181. https://doi.org/10.1111/cgf.14517

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *CoRR* abs/2307.01952 (2023). https://doi.org/10.48550/ARXIV.2307.01952

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. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, Vol. 139. PMLR, 8748–8763.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*. IEEE/CVF, 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, Vol. 9351. Springer, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

rqdwdw. 2023. negativeXL. https://civitai.com/models/118418/negativexl. Accessed: DATE 2023-02-10.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *CVPR*. IEEE/CVF, 22500–22510. https://doi.org/10.1109/CVPR52729.2023.02155

runwayml. 2024. stable-diffusion-v1-5. https://huggingface.co/runwayml/stable-diffusion-v1-5. Accessed: DATE 2024-01-02.

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. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. https://openreview.net/forum?id=M3Y74vmsMcY. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *ICML*, Vol. 37. JMLR.org, 2256–2265.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021a. Denoising Diffusion Implicit Models. In *ICLR*. OpenReview.net.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021b. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*. OpenReview.net.

Tsai-Ho Sun, Chien-Hsun Lai, Sai-Keung Wong, and Yu-Shuen Wang. 2019. Adversarial Colorization of Icons Based on Contour and Color Conditions. In *ACM MM*. ACM, 683–691. https://doi.org/10.1145/3343031.3351041

Daniel Sýkora, John Dingliana, and Steven Collins. 2009. LazyBrush: Flexible Painting Tool for Hand-drawn Cartoons. *Comput. Graph. Forum* 28, 2 (2009), 599–608. https://doi.org/10.1111/j.1467-8659.2009.01400.x

TencentARC. 2024. Hugging Face/IP-Adapter. https://github.com/TencentARC/T2I-Adapter/tree/SD. Accessed: DATE 2024-01-02.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In *CVPR*. IEEE/CVF, 1921–1930. https://doi.org/10.1109/CVPR52729.2023.00191

Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. 2024. InstantStyle: Free Lunch towards Style-Preserving in Text-to-Image Generation. *arXiv preprint arXiv:2404.02733* (2024).

Dingkun Yan, Ryogo Ito, Ryo Moriai, and Suguru Saito. 2023. Two-Step Training: Adjustable Sketch Colourization via Reference Image and Text Tag. *Computer Graphics Forum* (2023). https://doi.org/10.1111/cgf.14791

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *CoRR* abs/2308.06721 (2023). https://doi.org/10.48550/ARXIV.2308.06721

Yuno779. 2023. https://civitai.com/models/9409. Accessed: DATE 2023-06-25.

Lvmin Zhang. 2023. How ControlNet-reference works. https://github.com/Mikubill/sd-webui-controlnet/discussions/1236.

Lvmin Zhang. 2024. ControlNet-v1-1-nightly. https://github.com/lllyasviel/ControlNet-v1-1-nightly. Accessed: DATE 2024-01-02.

Lvmin Zhang, Chengze Li, Tien-Tsin Wong, Yi Ji, and Chunping Liu. 2018. Two-stage sketch colorization. *ACM Trans. Graph.* 37, 6 (2018), 261. https://doi.org/10.1145/3272127.3275090

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*. 3836–3847.

Richard Zhang, Phillip Isola, and Alexei A. Efros. 2016. Colorful Image Colorization. In *ECCV*, Vol. 9907. Springer, 649–666. https://doi.org/10.1007/978-3-319-46487-9_40

Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, and Alexei A. Efros. 2017. Real-time user-guided image colorization with learned deep priors. *ACM Trans. Graph.* 36, 4 (2017), 119:1–119:11. https://doi.org/10.1145/3072959.3073703

Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. 2023a. ProSpect: Prompt Spectrum for Attribute-Aware Personalization of Diffusion Models. *ACM Trans. Graph.* 42, 6 (2023), 244:1–244:14.

Yuechen Zhang, Jinbo Xing, Eric Lo, and Jiaya Jia. 2023c. Real-World Image Variation by Aligning Diffusion Inversion Chain. (2023).

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *ICCV*. IEEE/CVF, 2242–2251. https://doi.org/10.1109/ICCV.2017.244

Changqing Zou, Haoran Mo, Chengying Gao, Ruofei Du, and Hongbo Fu. 2019. Language-Based Colorization of Scene Sketches. *ACM Trans. Graph.* 38, 6 (2019). https://doi.org/10.1145/3355089.3356561

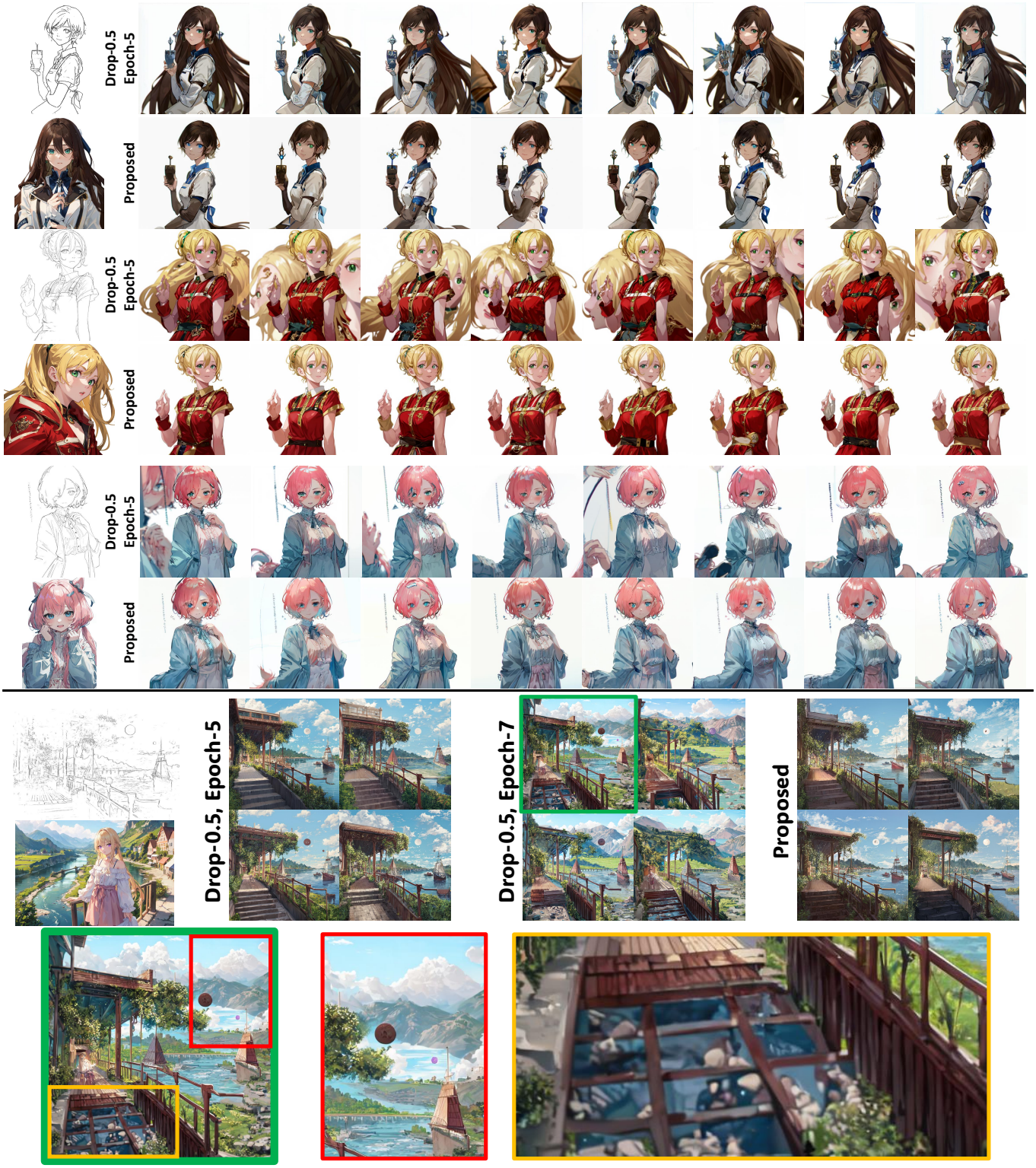

Fig. 10. Results generated in one batch with GS set to 5 and CAS to 1 by respective models. As seen in the upper comparison, the five-epoch *Drop-0.5* model shows a much higher probability of generating spatial entanglement compared to the proposed model. This tendency increases as training continues, highlighted in the bottom comparison, where compositions of results generated by the seven-epoch *Drop-0.5* model are visually chaotic. Sketches: ©SDAI(upper), ©tarotaro(lower).

Fig. 11. Qualitative comparison with baseline methods. Baseline results were generated using different sets of CASs that stress the transfer of style rather than composition, while ours used GS 5 and CASs 1 to demonstrate the improvement in diminishing the distribution problem. Baseline combinations are labeled as {Adapter of control condition}-{Control condition}-{Image prompt adapter}-{SD model} from top to bottom. Specifically, we fine-tuned *IP-Adapter v1.5* with *Anything v3* to align their distributions, labeled as *IP-Adapter-ft*. Sketches of (a),(c),(g) and references of row (c),(h): ©tarotaro; references of row (g): ©HuuOliv.