Enhancing Sketch Colorization via Separating Reference Representations

*Dingkun Yan, *Xinrui Wang†, Zhuoru Li, Jinze Yu, Suguru Saito, Yusuke Iwasawa, Yutaka Matsuo, Jiaxian Guo

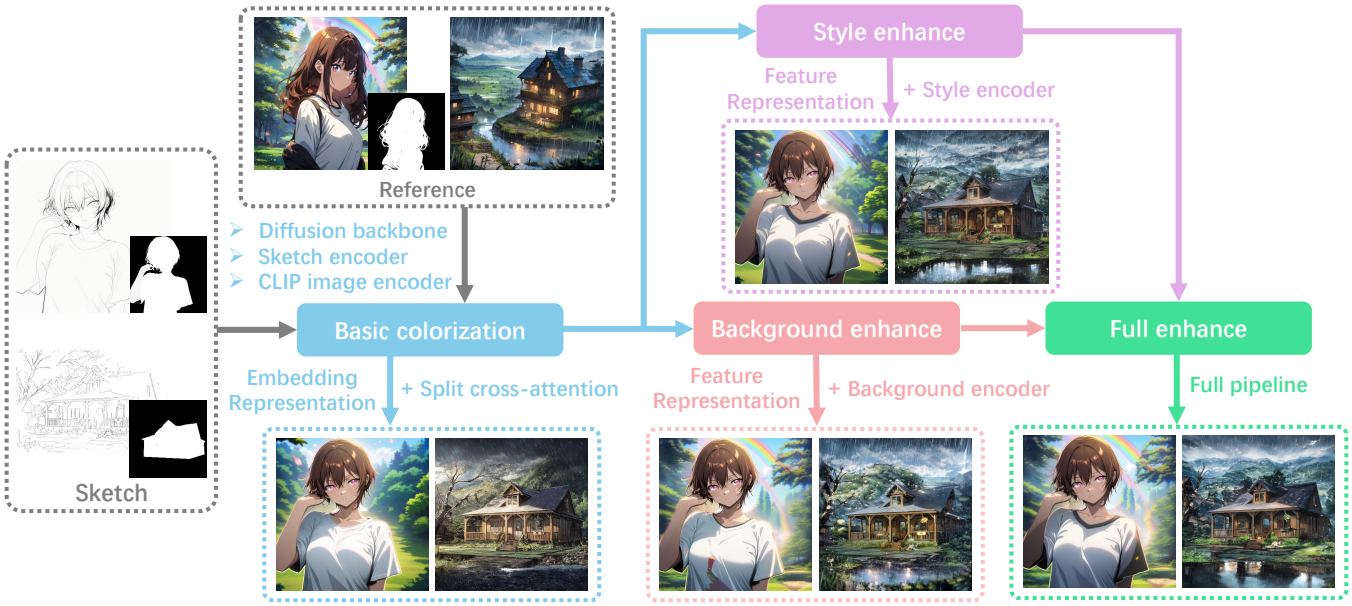

Fig. 1: We propose a reference-based sketch colorization workflow that employs separate reference representations with respective modules and is trained with a multi-stage schedule. The system achieves state-of-the-art anime-style colorization results without requiring spatial correspondence between inputs.

***Abstract*—Reference-based sketch colorization methods have garnered significant attention for the potential applications in animation and digital illustration production. However, most existing methods are trained with image triplets of sketch, reference, and ground truth that are semantically and spatially similar, while real-world references and sketches often exhibit substantial misalignment. This mismatch in data distribution between training and inference leads to overfitting, consequently resulting in spatial artifacts and significant quality degradation in colorization results. To address this issue, we carefully analyzed reference representations, defined as the intermedium to transfer information from reference to sketches. Building on our findings, we present a novel framework that leverages distinct reference representations to separately colorize different regions of the sketches in multiple stages, enhancing visual quality and reference similarity, while mitigating spatial artifacts. Specifically, we follow the real-world animation production workflows to introduce a split cross-attention mechanism that separately processes the foreground regions and background regions. A backbone network guided by high-level semantic embeddings is trained in the first stage for coarse colorization, a background encoder and a style encoder are then trained in separate stages to enhance low-level feature transfer and improve reference similarity. This design also enables flexible inference modes suitable for various use cases. Extensive qualitative and quantitative evaluations, together with user studies, demonstrate the superior performance of our proposed method compared to existing approaches. Code and pre-trained weight will be made publicly available upon paper acceptance.**

***Index Terms*—sketch colorization, diffusion models**

*Represent equal contribution, †Represent corresponding author, Email address: secret_wang@weblab.t.u-tokyo.ac.jp

Dingkun Yan and Suguru Saito are with Institute of Science Tokyo, Tokyo, 152-8550, Japan.

Xinrui Wang, Yusuke Iwasawa, Yutaka Matsuo and Jiaxian Guo are with The University of Tokyo, Tokyo, 113-8654, Japan.

Zhuoru Li is with Project HAT, Xiamen, 351000, China

Jinze Yu is with Department of Communications and Computer Engineering, Waseda University, 169-8555, Tokyo, Japan

I. Introduction

Animation has long been a popular artistic form and been in great demand by the audience around the world for decades, and increasing market demand is straining the capacity of animation studios, posing significant challenges to the industry. Within current animation creation workflows, sketch colorization represents a particularly labor-intensive process, occupying a substantial portion of studio personnel. Consequently, machine learning techniques have been explored to automate this task and alleviate manual effort.

Early attempts utilizing Generative Adversarial Networks (GANs) [2]–[5] yield suboptimal colorization results due to the limitation in generative capacity. Diffusion models have recently been applied to sketch colorization because of their ability to synthesize high-quality images, and methods with images as references are the most popular. Based on training design, they can be categorized into two types: 1. using reference inputs derived from the ground truth [6]–[8]; 2. using

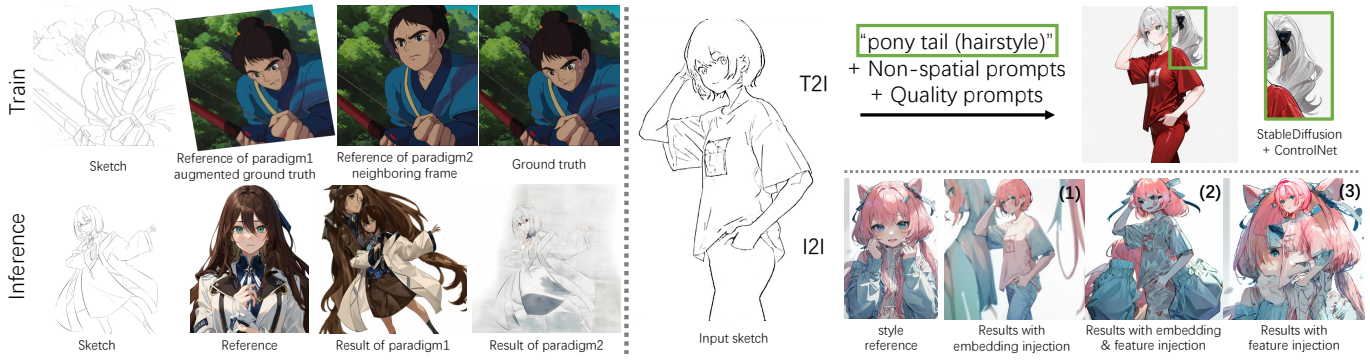

Fig. 2: Left: data used for training and inference exhibit a significant gap in the semantics and geometry between reference images and sketches due to their different sources. This gap results in severe deterioration in the colorization quality. Right: spatial entanglement caused by this gap when adopting the first training paradigm, which is termed as distribution shift in [1]. This issue makes the T2I method mistakenly change the hairstyle in the result, and artifacts in the I2I methods are influenced by the intermedium of reference injection. The artifacts increase when there is more detailed information injected into the common training scheme. Training frames come from movie *Princess Mononoke*

images with the same contents as reference, such as neighboring frames in a video clip [9] or manga grids [10] and directly transfer low-level representations during training. Shown in the left panel of Figure 2, both training paradigms use reference images similar to ground truth as training data, introducing distribution shift during inference, as reference images and sketches are usually spatially and semantically less relevant for inference. The first kind of methods usually experience spatial artifacts called spatial entanglement [6], showing as additional characters or body parts in the background region or unexpected changes of sketch semantics, while the second kind of methods tend to synthesize insufficiently colorized results with blurry details.

In this paper, we follow the first paradigm to use images derived from the ground truth as references, and propose a multi-stage colorization framework to eliminate spatial entanglements and enhance colorization qualities. We start from analyzing the spatial entanglements caused by distribution shift, which exhibit as artifacts and severe deterioration in image quality for inputs with misaligned semantics or structures. This occurs because the intermediums used to inject reference information into diffusion models during training, termed ***reference representations*** in this paper, contain not only color and style information used for colorization, but also spatial content that might contradict with the sketch semantics. When injected into the diffusion backbone, the misaligned content information causes incorrectly colorized regions beyond the control of the sketches. We further discovered that different reference representations used for reference injection are causing corresponding distinct artifacts. Specifically, embeddings with higher-level semantic information as reference representations tend to produce results with fewer artifacts but blurry textures, while latent features with richer details and lower-level semantics as reference representations yields results with higher similarity but more pronounced artifacts. The above mentioned different kinds of artifacts are illustrated with corresponding reference representations in the right panel Figure 2.

Inspired by real-world animation production workflows (shown in Figure3) where foreground characters and backgrounds are colorized separately, we design a split cross-attention mechanism [1] with spatial masks to segment the foreground and background regions in both the sketch and reference images, and trainable LoRA (low rank adaption) [11] to independently process the two regions, ensuring faithful color transfer while preventing interference and spatial entanglement. However, this mechanism, when combined with embedding reference representation, degrades the texture and detail transfer and cause blurred colorization results. Therefore, we introduce a background encoder collaborated with feature representations to facilitate the transfer of fine-grained details from the reference image, especially the background regions. Furthermore, a style encoder is introduced to better transfer style information such as tone and textures to the full image. To preserve the network's understanding of the spatial semantics of sketch images and prevent different reference representations used in different modules from influencing each other, we design a multi-stage training strategy that separately trains each component. This design also enables the background encoder and style encoder to learn the residuals between the ground truth and results of previous stage, greatly simplifying the optimization.

We conduct extensive experiments, including qualitative evaluations showing that our method generates high-quality results and faithfully transfers the color and textures from reference images while avoiding spatial entanglement, quantitative comparisons against existing methods validating the superiority of the proposed approach, and a comprehensive ablation study demonstrating the contribution of each component in mitigating spatial entanglement and artifacts in various scenarios. Finally, user studies shows that users prefer our method over existing methods.

In summary, our contributions are as follows: (1) We provide a detailed analysis of reference-based sketch image colorization methods and identify the underlying causes and manifestations of artifacts. (2) We propose a novel coloriza-

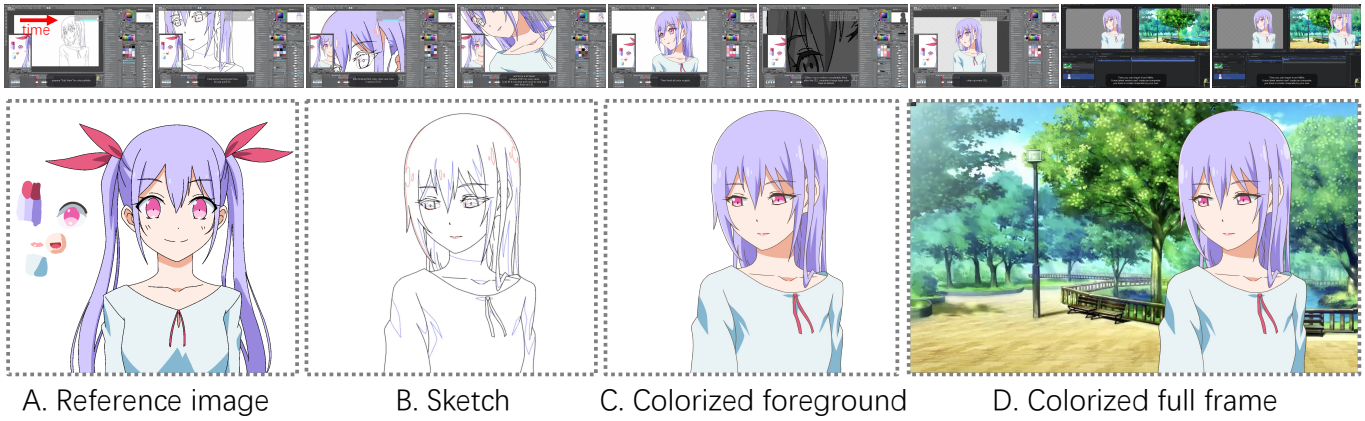

Fig. 3: Illustration of colorization workflow in professional animation studios. A: character designers design characters as references. B: Senior animators draw the sketches for the key frames. C: animators colorize the figures in the sketches according to the character designs, and D: animators colorize the background of the sketches and merge foreground and background into finished frames.

tion framework with components designed to address specific artifact types, resulting in effective mitigation and high-quality colorization without requiring well-aligned input pairs. (3) Extensive experiments demonstrate the superiority of our method over existing approaches through qualitative and quantitative comparisons, as well as a perceptive user study.

II. Related Work

A. Latent Diffusion Models

Diffusion Probabilistic Models [12], [13] are a class of latent variable models inspired by nonequilibrium thermodynamics [14] and have achieved great success in image synthesis and editing. Compared to Generative Adversarial Networks (GANs) [15]–[17], Diffusion Models excel at generating high-quality images with various contexts and stronger control over different conditional guidance. However, the autoregressive denoising process of diffusion models, typically computed with a U-Net [18] or a Diffusion Transformer (DiT) [19], [20], incurs substantial computational costs. To reduce this cost, Rombach et al. proposed Stable Diffusion (SD) [21], [22], a class of Latent Diffusion Models (LDMs) that performs a denoising process in a perceptually compressed latent space with a pair of pre-trained Variational Autoencoder (VAE). Also, studies on accelerating the denoising process have demonstrated effectiveness [13], [23]–[25]. In this paper, we adopt SD as the backbone, utilize the DPM++ solver [13], [25], [26] as the default sampler, and employ classifier-free guidance [27], [28] to strengthen the transfer performance.

B. Image Prompted Diffusion Models

Existing diffusion based methods have achieved notable progress in text-guided generation [19]–[22]. However, many tasks require more detailed guidance that provides better control on the generated content, including image-to-image translation [29], style transfer [30], [31], colorization [32], [33] and image composition [34], [35]. In these tasks, images are used as prompts, and the reference information extracted from prompt images varies from task to task: style transfer prefers textures and colors, image composition focuses more on object-related information, and sketch colorization requires all above.

Two primary intermediums are commonly used to transfer reference information to diffusion models: image embeddings extracted by pre-trained vision encoders [5], [8], [33], [36] and latent features directly encoded from reference images by jointly trained modules [7], [32], [37]–[40]. However, these intermediums, termed as reference representation in this paper, may introduce mismatched structural and semantic information, particularly when the input sketch and reference image are poorly aligned, leading to degraded generation quality. In the context of sketch colorization, these reference representation unavoidably provide conflicting spatial cues during inference, resulting in unacceptable artifacts as depicted in Figure 2.

C. Sketch Colorization

Machine learning methods have been widely adopted to ACG related tasks such as cartoon [41], [42] and manga [43] generation. Among all test topics, sketch colorization has been the most focused. Early approaches started from interactive optimization [44], deep learning then emerged as the dominant paradigm, enabling the synthesis of high-quality, high-resolution color images [3], [33], [40], [45]. Based on the guiding modality, existing methods can be broadly categorized into three groups: text-prompted [5], [45], [46], user-guided [3], [47], and reference-based [5], [32], [40]. User-guided methods offer fine-grained control through direct user input (e.g., spots, sprays), but their reliance on manual intervention hinders their applicability in automated pipelines. Text-prompted methods, driven by advancements in text-to-image diffusion models, have gained significant traction, yet it is challenging to precisely control colors, textures, and styles using text prompts.

Image-referenced methods have also benefited from advancements in diffusion models and image control techniques [7], [8], [46], [48], [49]. However, a critical challenge remains: effectively addressing the spatial and semantic conflicts

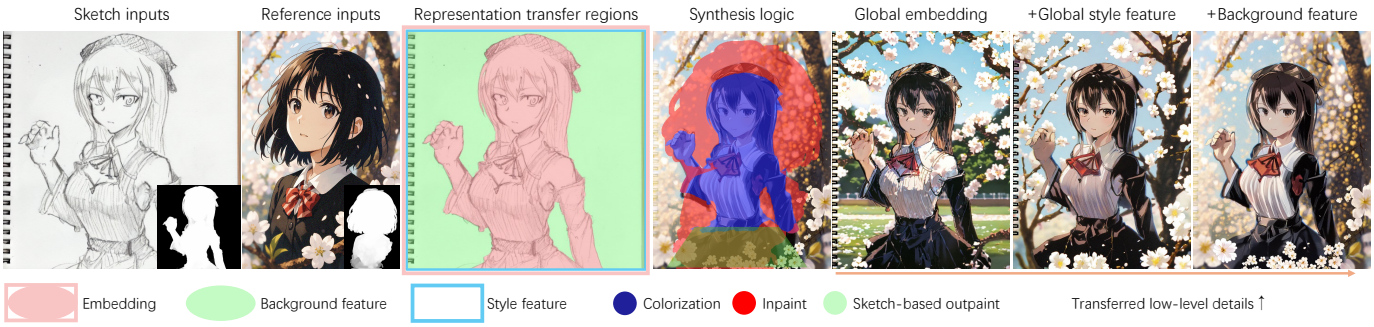

Fig. 4: Illustration of the proposed reference-based sketch colorization workflow. To eliminate artifacts and enhance colorization quality, we separate colorization into three parts, leveraging foreground masks extracted from the reference and sketch inputs: embedding guidance for sketch-covered regions, style modification for global details, and low-level transfer for non-sketch regions. Moreover, the network should be able to properly inpaint the missing regions based on neighboring features in the sketch and reference images. As highlighted by red, the proposed network inpaints the skirt based on prior knowledge from the sketch and the flowers based on neighboring features from the reference.

between diverse reference images with the often sparse and abstract nature of sketches during inference. Existing methods typically require extracted sketches with highly-matched references as input pairs to achieve satisfying results [5], [32], [38]–[40], [50], limiting their generalizability and potential applications. While ColorizeDiffusion [33] demonstrated significant progress in colorization quality, it still grapples with spatial entanglement, as visualized in Figure 2, and struggles to accurately transfer fine-grained details. In this paper, we address these limitations by introducing a novel step-by-step training strategy within a refined colorization framework designed to explicitly mitigate spatial entanglement.

III. Method

A. Overview

In this paper, we propose a multi-stage sketch colorization framework with separated reference representation to reduce the spatial entanglements and enhance the stylization quality. The framework consists of a diffusion backbone, a variational autoencoder (VAE), a sketch encoder, a CLIP image encoder, a style encoder, and a background encoder. It leverages a sketch image $X_s \in R^{w_s \times h_s \times 1}$, a reference image $X_r \in R^{w_r \times h_r \times c}$, a sketch mask $X_{sm} \in R^{w_s \times h_s \times 1}$ and a reference mask $X_{rm} \in R^{w_r \times h_r \times 1}$ as inputs, and returns the colorized result $Y \in R^{w_s \times h_s \times c}$, with w, h and c representing the width, height and channel of the images.

Following the real-world animation production pipeline, we design a split cross-attention mechanism [1] that separately processes the foreground and background regions of both sketch and color references with spatial masks, respectively, and integrate it into the diffusion backbone, the background encoder and the style encoder. We further propose a multi-stage colorization framework that optimize different modules with corresponding reference representation. In the first stage, we jointly train the colorization backbone and a sketch encoder. In the second stage, the backbone and sketch encoder trained in the first stage are fixed, and a background encoder is trained with feature representation as references to synthesize the finer details in the background region. In the third stage, we fix all the modules trained in previous stages, and optimize a Style encoder with feature representation as references to enhance the style information transfer such as tone and textures to the full image. Furthermore, we design a character-mask merging strategy together with a background bleaching strategy to inpaint the possible overlap regions of reference masks and sketch masks, making the synthesis of boundary regions natural. The colorization workflow and multi-stage training pipeline are visualized in Figure 4 and Figure 5, respectively. The details of character-mask merging and background bleaching are included in the supplementary material.

B. Analysis and Extraction of Reference Representations

Reference representations used for sketch colorization can be categorized by semantic level from high to low into three types: 1. discrete embeddings, (e.g., CLIP text embeddings) [46], 2. continuous embeddings (e.g., CLIP image embeddings) [7], [33], and 3. Image features (e.g., extracted feature maps) [38], [39]. As illustrated in the right panel of Figure 2, Each of the three reference representations is sub-optimal to transfer reference information to colorization backbone when used alone, and all of them may introduce spatial entanglements. Especially, text embeddings lack spatial information and fail to apply precise spatial and color control on the colorization results; Image embeddings contain high-level spatial and semantic information. It is the most suitable representation among all three when used alone, but may fall short in transferring details to the synthesized result; Image features contain rich low-level information which helps reconstruct textures and details, but excessive low-level semantics worsen the spatial entanglements. Based on this analysis, we propose to employ both the CLIP image embedding and extracted feature maps as reference representation to guarantee high-quality colorization with fine details, and design a split cross-attention mechanism together with a multi-stage training framework to mitigate the spatial entanglements. We illustrate the framework in Figure 5.

We exploit the OpenCLIP-H image encoder [51]–[53] to extract image embeddings as the embedding representations.

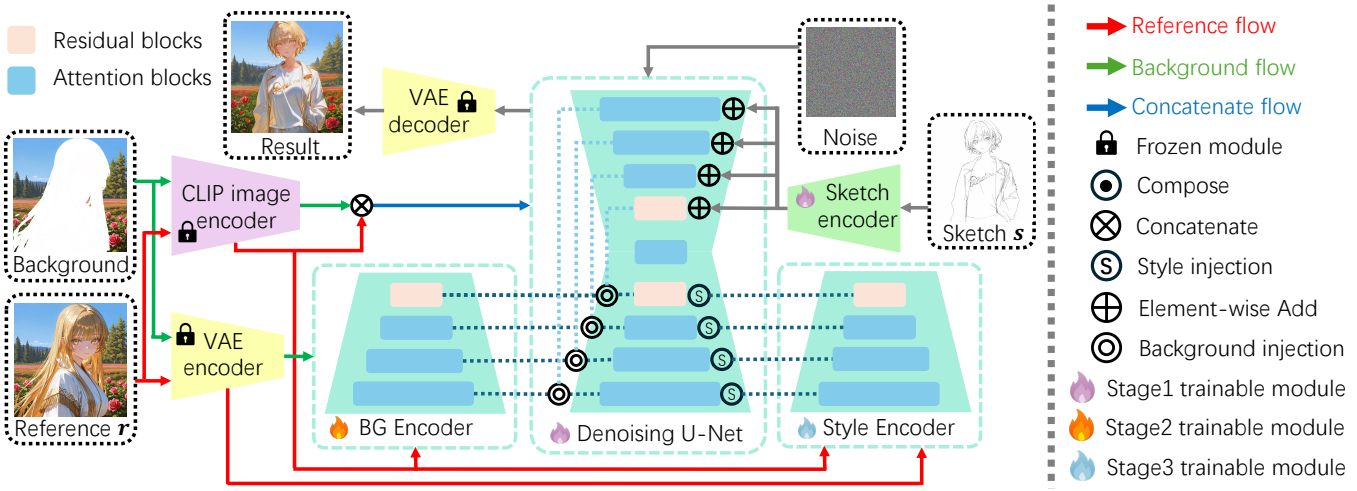

Fig. 5: Illustration of the proposed framework. The CLIP image encoder and the VAE encoder are fixed during training. The extracted image embeddings and latent images are injected into the corresponding modules in the same way as standard LDM. The denoising U-Net, the background encoder and the style encoder are trained separately in 3 stages. Detailed pipelines of stage 1 and stage 2 are included in the supplementary materials.

The pre-trained ViT-based image encoder network extracts 2 kinds of image embeddings: the CLS embeddings $E_{cls} \in R^{bs\times1\times1024}$ and the local embeddings $E_{local} \in R^{bs\times256\times1024}$, where bs represent the batch size. Previous image-guided diffusion models [7], [30] commonly utilize CLS embeddings as color or style references, which are connected to text-level notions and projected to CLIP embedding space for image-text contrastive learning, with spatial information compressed as spatial semantics. ColorizeDiffusion [33], on the other hand, reveals that local embeddings also express text-level semantics and, meanwhile, express more spatial details regarding textures, strokes, and styles, enabling the network to generate better reference-based results with finer details. Consequently, the proposed method follows [33] to adopt local embeddings as a reference representation. The extracted embedding representations are then injected into the diffusion U-Net through split cross-attention.

Nevertheless, the inherent lack of low-level, detailed reference information in image embeddings can lead to blurry textures in the colorized output. For sketches depicting isolated figures or featuring simple background lines, the information encoded within image embeddings may be insufficient to reconstruct a meaningful and visually compelling background. To address this limitation, we incorporate a style encoder and a background encoder to extract latent features as a reference representation to facilitate the reference information injection. The background encoder and the style encoder are initialized from the encoder of diffusion U-Net trained in stage 1 and optimized in stage 2 and stage 3, respectively, and the features extracted by them are injected into the diffusion backbone through corresponding injection modules as reference representations.

C. Split Cross Attention

In anime images, the foreground regions and background regions differ distinctively in color distribution, color block sizes, tones, and textures. Thus, the colorization of foreground and background is separated into two independent steps in the animation production workflow. Following this scheme, we propose a novel split cross-attention mechanism to substitute the cross-attention layers in the diffusion backbone to separately process foreground and background regions with different parameters in a single forward pass. The architecture of the split cross-attention is illustrated in Figure 7.

A split cross-attention layer consists of two groups of trainable LoRA weights W_f^t and W_b^t, which include query weights W_f^q and W_b^q, key weights W_f^k and W_b^k, and value weights W_f^v and W_b^v for foreground and background QKV projection respectively. An open-sourced animation image segmentation tool [54], [55] is used to automatically extract the foreground mask m_s and m_r of sketches and reference images. Regions with pixel values larger than thresholds ts_s and ts_r are considered as foreground, otherwise background. Following [56], we set the ranks of foreground LoRAs as 16; for background LoRA, the rank is formulated as $r = 0.5 * min(D_q, D_{kv})$, where D_q and D_{kv} are dimensions of queries and keys/values for the corresponding cross-attention layers. Especially, for the reference images, only characters are regarded as foreground, otherwise the whole image is regarded as background. For the sketches, however, the foreground is defined as salient objects and all images are separately processed. We also include the discussion about mask segmentation in the supplementary materials.

We define query inputs (forward features) as z_f, z_b, key and value inputs (reference embeddings) as E, E_b, attention outputs as y in the following sections, where the index f and b indicate foreground and background respectively. Specifically, E denotes the reference embeddings extracted from the whole reference image r, formulated as $E = \phi(X_r)$; and $E_b = \phi(X_{rb})$, where X_{rb} is the background region of the reference image, and ϕ represent the feature extraction network. During training, the proposed split cross-attention can be formulated as follows:

$$y = \begin{cases} \textbf{Softmax}(\frac{(\hat{W}_f^q z_f)\cdot(\hat{W}_f^k E)}{d})(\hat{W}_f^v E) & \text{if } m_s > ts_s \\ \textbf{Softmax}(\frac{(\hat{W}_b^q z_b)\cdot(\hat{W}_b^k E_b)}{d})(\hat{W}_b^v E_b) & \text{if } m_s \le ts_s \end{cases} \quad (1)$$

where $\hat{W}_f^t = W^t + W_f^t$, and W^t represents the pre-trained weights, which remain frozen during training. Similarly, $\hat{W}_b^t$ follows the same approach.

D. Feature Injection Modules

In existing colorization methods, spatial entanglements are mainly represented as foreground exceeding boundaries and appearing in the background, such as extra characters or body parts. An efficient way to eliminate the phenomena is to separately process the foreground and background with a split cross-attention mechanism [57], where sketch-guided regions are regarded as foreground and all other regions are regarded as background. However, utilizing this mechanism usually degrades the quality and textures of colorization results. To further enhance textures and semantics synthesis, particularly for sketches with sparse or absent background lines, we introduce a background encoder and a style encoder to transfer features with detailed information to backgrounds and facilitate the synthesis of fine textures.

As illustrated in Figure 8, we design a background injection module and a style injection module to integrate the output of background encoder and style encoder respectively into the denoising U-Net decoder at corresponding levels. We denote the sketch mask as m_s, the user-defined foreground threshold for the sketch input as ts_s, computation of the transformer block as $\mathcal{W}(\cdot)$, the skip feature from the encoder of the denoising U-Net as z_{skip}, and the feature from the background encoder as z_{bg}, the calculation of the background injection module is formulated as:

$$z_{skip} = \begin{cases} z_{skip} & \text{if } m_s > ts_s \\ \mathcal{W}(z_{skip}, z_{bg}) & \text{if } m_s \le ts_s. \end{cases} \quad (2)$$

For the style injection module, we adopt adaptive normalization [58] to enhance global style transfer. Given the forward feature in denoising U-Net as z, the style feature from the style encoder as z_{style}, global average pooling as GAP$(\cdot)$, timestep embedding as E_t, the style modulation $\mathcal{M}(\cdot)$ can be formulated as

$$\mathcal{M}(z, \hat{z}_{scale}, \hat{z}_{shift}, E_t) = z \cdot (1 + \hat{z}_{scale}) + \hat{z}_{shift}, \quad (3)$$

where $\hat{z}_{scale}$ and $\hat{z}_{shift}$ are obtained via linear projections from z_{style}, conditioned on the timestep embedding E_t.

E. Multi-step Training Strategies

The proposed framework consists of multiple components, with each component employing different reference representations and responsible for the colorization of different regions and granularities. When jointly trained, the mixed reference representations and various learning objectives of different modules may hinder the optimization of each other. To tackle this issue, we further design a multi-stage training strategy to train the diffusion backbone, the background encoder, and the style encoder separately.

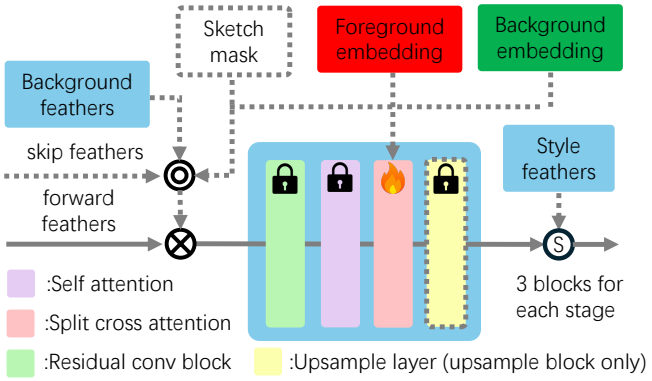

Fig. 6: Illustration of the detailed architectures of the U-Net blocks in Diffusion backbone.

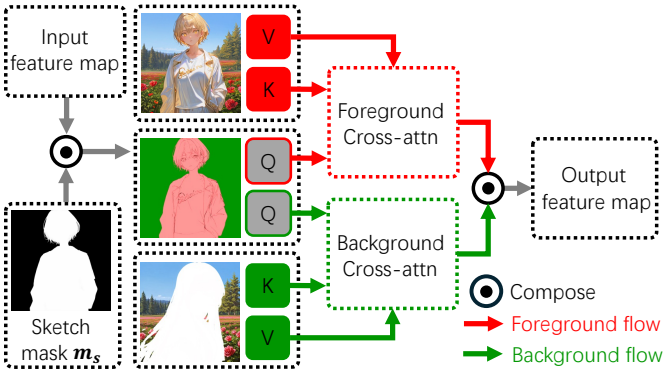

Fig. 7: Illustration of the detailed architectures of the split cross attention.

1. Colorization pre-training stage: we only optimize the sketch encoder and denoising U-Net in this stage. We at first follow [33] to train the diffusion backbone with a two-stage noisy training-refinement scheduler, and then follow [1] to optimize the split cross-attention module added into the diffusion backbone. A dynamic reference drop of 80% for the noisy training stage and 50% for the refinement stage is adopted to avoid severe deterioration in the segmentation and perceptual quality of results.

2. Foreground-background separated training stage: we add a background encoder into the framework and optimize it together with the split cross-attention module, with other parameters frozen. This stage helps eliminate spatial entanglement caused by the reference embeddings and enhances the synthesis of backgrounds;

3. Hybrid training stage of style encoder: in this stage, we add a style encoder into the framework and optimize it with the parameters in diffusion backbone and sketch encoder frozen. The parameters in background encoder and split cross-attention modules are not trained but randomly activated at a rate of 50% to generate extra conditions for the denoising backbone and other parameters frozen.

In stage 2 and stage 3, the reference embeddings for denoising U-Net are dropped at a fixed rate of 50%. Given noise

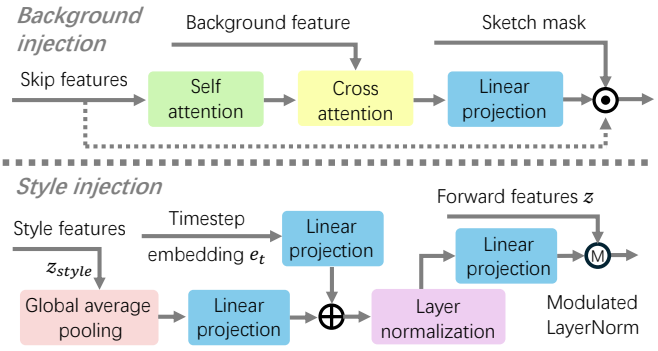

Fig. 8: Illustration of the proposed feature injection modules.

ϵ, sketch s, ground truth y, encoded latent representations (forward features) z_t at timestep t, VAE encoder $\mathcal{E}$, denoising U-Net and sketch encoder θ, background encoder with background injection φ_{bg}, style encoder with style injection φ_{style}, and CLIP image encoder ϕ, the training objective for all training stages can be defined as

$$\mathcal{L}(\vartheta) = E_{\mathcal{E}(y),\epsilon,t,s,c}[\|\epsilon - \epsilon_\theta(z_t, t, s, c)\|_2^2], \tag{4}$$

where ϑ and c represent the optimization targets and conditional inputs, and a detailed explanation for each stage is as follows. **Stage 1**: ϑ represents the denoising U-Net and the sketch encoder, and c represents image embeddings $e = \phi(r)$; **Stage 2**: ϑ represents the background encoder and injection modules, as well as LoRAs inside split cross-attention layers, and c represents background embeddings e_{bg}, sketch mask m_s, and background features $z_{bg} = \varphi_{bg}\big(\mathcal{E}(r_{bg}), e\big)$; **Stage 3**: ϑ represents the style encoder and injection modules, and c represents style features $z_{style} = \varphi_{style}\big(\mathcal{E}(r), e\big)$. As ground truths are directly used as references, r can be replaced with y throughout all equations.

Specifically, the style encoder is trained in Stage 3, following the training of the background encoder. This scheduling helps prevent the style encoder from capturing excessive content information.

F. Switchable Inference Mode

To accommodate diverse application scenarios and control requirements, the proposed framework integrates three distinct inference schemes, selectively engaging specific architectural components: the *Vanilla* mode, the *Style enhance* mode, and the *Background enhance* mode.

Vanilla Mode is introduced for two specific scenarios: (1) the color reference exhibits low resolution or quality, limiting the value of detailed feature representations, and (2) the input sketch itself contains highly intricate compositions, such as landscapes, and provides sufficient spatial information for reconstruction. In both cases, the primary objective of the Vanilla Mode is the faithful reconstruction and colorization of the input sketch's semantic structure, prioritizing structural integrity over high-frequency style transfer.

Style enhancement mode activates the style encoder during the denoising process, enabling the effective transfer of fine-grained textures and brushstrokes. While this mode represents a stylistic extension of the Vanilla Mode, its intensified focus on high-frequency texture fidelity may, in specific corner cases, introduce a slight degradation in the semantic segmentation consistency of the final output.

Background enhancement mode functions as a mask-guided approach. By activating both the background encoder and the split cross-attention layers, it enforces an explicit feature separation between the foreground and background streams throughout the forward pass. This mechanism effectively mitigates spatial entanglement and is particularly effective in generating detailed, high-fidelity backgrounds with superior texture and fine details.

Full enhancement mode simultaneously activate both style enhancement and background enhancement, which yields optimal, comprehensive transfer performance across both style and background fidelity. A detailed comparative analysis of these schemes and their respective performance metrics is provided in Section 5.

IV. Experiment

A. Implementation

Dataset, configurations, and environment. We curated a dataset of over 6M (sketch, color, mask) image triples from Danbooru [59], encompassing a diverse range of anime styles for illustration-format images. Sketches were generated by jointly using [60] and [61], while masks were produced using [54]. The dataset was divided into a training set and a validation set of 52,000 triples without overlap. The training was conducted on 8x H100 (94GB) GPUs utilizing Deepspeed ZeRO2 [62] and the AdamW optimizer [63], [64] with a total batch size of 256, a learning rate of 0.00001, and betas of (0.9, 0.999). During training, images are resized and cropped to random sizes between 768 to 1280 and random aspect ratios across batches, while all images used for validation are resized to 512*512. More details are described in the supplementary material.

Compared baseline. We compare our method with existing reference-based sketch image colorization methods Yan et al. [5], MangaNinja [9], ColorizeDiffusion [1], IP-Adaptorip-adapter,controlnet-iccv, T2I-Adaptor [8], Cobra [10] to demonstrate the superiority of the proposed framework. All results of previous methods used for comparison are generated by their official implementations. Depending on training schemes introduced in Section I, baseline image-guided methods can be categorized into two classes: 1. training with augmented ground truth as references [1], [5], [7], [8]; 2. training with references that contain the same identities [9], [10]. Given the complexity of adapter-based baselines, we provide a concise overview of their operation and integration into the image-guided sketch colorization in the supplementary materials.

B. Qualitative Comparison.

We present two qualitative comparisons between our proposed method and existing methods in Figure 10 and Figure 9. Among the baselines, MangaNinja [9] and Cobra [10] are

Fig. 9: Qualitative comparisons regarding figure colorization. Different from recent colorization baselines [1], [9], [10], the proposed method is demonstrated to be superior in the quality and similarity of colorization without having spatial entanglement and requiring inputs to have semantically or spatially similarity. High-resolution images and user hints used for Cobra [10] are contained in the supplementary materials.

specifically designed for character colorization. These methods mitigate spatial entanglement artifacts by using training references composed of different images that feature the same identities. However, this strategy often leads to overfitting, resulting in notable performance degradation when the input sketch and reference image are semantically or geometrically misaligned. This limitation reduces the generalization ability of their models. Consequently, these models exhibit limited generalization. Furthermore, we exhibit cross-content colorization of figure and landscape sketches in Figure 10. This scenario lacks clear correspondence between input identities, so subjective evaluations usually prioritize the similarity of style, color scheme, and texture/stroke details, all of which require a reasonable transfer of low-level visual features.

Especially, **IP-Adapter** and **InstantStyle** cause artifacts on the roof of F column in Fig 10 and obvious spatial

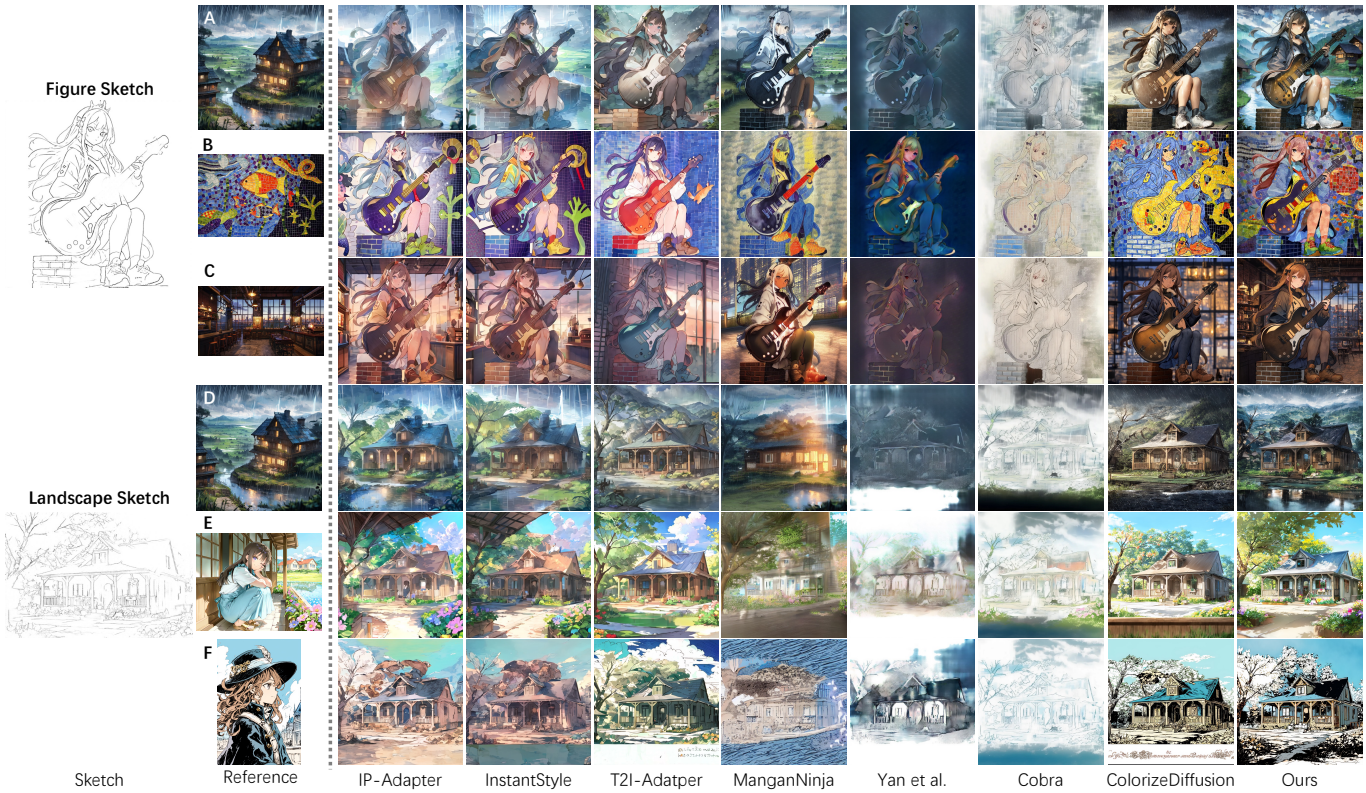

Fig. 10: Cross-content transfer results. Raw A-C show results of figure sketches, where ours significantly outperform in segmentation and colorization. Raw D-F show results of landscape sketches, where ours demonstrate better composition and transfer performance, as well as outpainting for non-sketch regions.

TABLE I: Quantitative comparison on 768^2 resolution between the proposed model and baseline methods. †: These evaluations randomly selected color images as references, making them close to real-application scenarios. ‡: Ground truth color images were deformed to obtain semantically paired and spatially similar references for evaluations. §: Tested at 512^2 resolution.

Method	†Aesthetic ↑	†FID ↓	‡PSNR↑	‡MS-SSIM↑	‡CLIP similarity↑
Ours	**5.1859**	**5.6330**	29.3626	**0.7081**	**0.9056**
ColorizeDiffusion	4.8351	9.6423	28.7215	0.5899	0.8753
IP-Adapter	4.6627	38.9232	28.5124	0.5464	0.8632
InstantStyle	4.7150	40.2134	28.0921	0.4467	0.8039
T2I-Adapter	4.2647	41.1569	28.1321	0.3194	0.7134
§*MangaNinja*	4.1738	42.9741	**29.5741**	0.6715	0.7304
§Yan et al.	4.7923	27.0032	29.1293	0.5239	0.8894

entanglement in column e, f in Fig 9. **T2I-Adapter** causes strange color tone in column F of Fig 10, and fails to correctly colorize the coat of column B and the trousers in column I of Fig 9. **MangaNinja** synthesizes results with sub-optimal overall color tone on column B, E and F of Fig 10 and column A, D, E and G of Fig 9. **Yan et al** performs poorly on complicate sketches, leading to results with palette degraded to a few colors in Fig 10 and low saturation results in Fig 9. **Cobra** struggles in cross domain colorization and collapses in Fig 10, and also generates poor results in columns A and B of Fig 9. **ColorizeDiffusion** functions well in use cases with complicated sketches but still suffers from spatial entanglement in columns E and F of 9. **The proposed method**, in contrast, loyally preserves the color distribution and tune of reference images in Fig 10 and synthesizes images with clear and high quality background and free of artifacts.

C. Quantitative Comparison.

In this paper, we evaluate the quantitative performance of the proposed method and existing baselines with 5 different criteria: Aesthetic score [65], Fréchet Inception Distance (FID) [66], multi-scale structural similarity index measure (MS-SSIM), peak signal-to-noise ratio (PSNR), and CLIP score [51]. Aesthetic score is a quantitative metric derived from human preferences, used to predict and guide the generation of images with high subjective quality and visual appeal based on factors like composition and style. FID is a standard metric for evaluating generative models, quantifying the similarity between different distributions of real and generated image features extracted by a pre-trained Inception network. Both metrics don't require inputs to be semantically and spatially paired. We conducted two evaluations using these metrics on the entire validation set, which contains 52k (sketch, reference) image pairs. Reference images were randomly selected during

Fig. 11: The separate processing effectively eliminates spatial entanglement artifacts in background regions, while the background encoder further enhances background synthesis. However, jointly training these components with the denoising U-Net negates these improvements by preventing the necessary disentanglement.

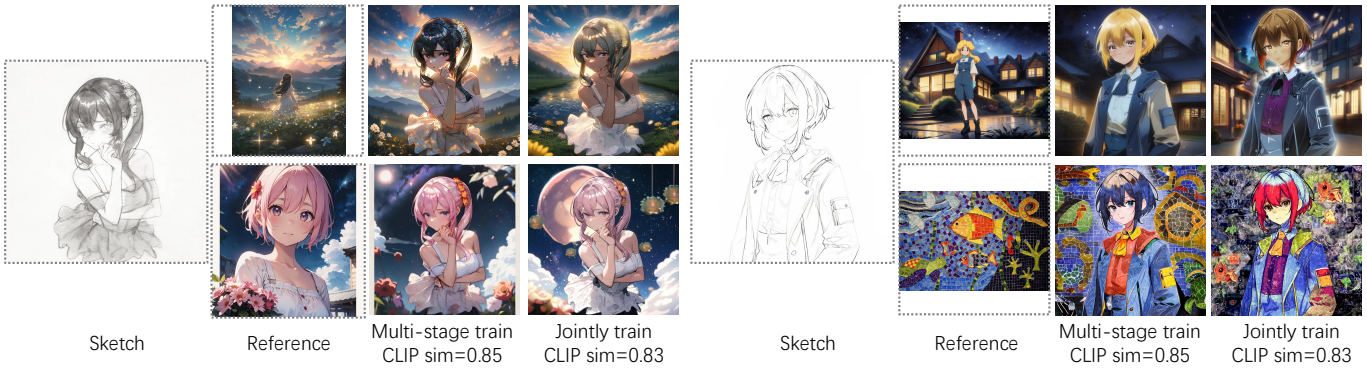

Fig. 12: Ablation study regarding training strategy for the style injection. We calculated 5K CLIP similarity to evaluate their transfer performance quantitatively in this comparison, and all results were generated using *style enhance* mode.

validation.

MS-SSIM assesses the similarity between two images by aggregating structural comparisons across multiple spatial resolutions. PSNR quantifies image reconstruction quality by calculating the ratio, in decibels, between the maximum possible signal power and the Mean Squared Error (MSE). Both MS-SSIM and PSNR are fully-referenced perceptual metrics. CLIP Score is a reference-free metric that evaluates the semantic alignment between a generated image and a text prompt (caption) by computing the cosine similarity between their respective visual and textual embeddings within the shared latent space of the pre-trained CLIP model. As these metrics require the reference to be aligned with ground truth, we selected 5000 color images as ground truth to generate extracted sketches and deformed references, where references were deformed using thin plate spline (TPS) transformation.

We show the results of quantitative evaluation in Table I. Due to the lack of a batch inference script in Cobra official implementation [10], we excluded it in this comparison. The proposed method significantly outperforms in all evaluations owing to the removal of artifacts, higher fidelity to the sketch composition, and stronger style transfer ability.

D. Ablation Study

Artifacts removal. We first validate the effectiveness of the proposed framework and training strategy in eliminating background artifacts caused by spatial entanglement [1], [6]. We set up four ablation frameworks: A) *vanilla* mode, B) *vanilla* mode with split cross-attention and trainable LoRAs, C) *style enhance* mode, D) *Full enhance* mode, and D) ablation result if related components are jointly optimized in stage 1.

We show a qualitative comparison in Figure 11 to validate the effectiveness of the proposed modules. The vanilla mode generates images in (A) with obvious artifacts in the background such as cake in the upper row and hairs in the lower row. Split cross-attention reveals the spatial entanglements in the background in (B), but there are still artifacts in the leg and sleeve regions in the upper row and unwanted stuff in the background in the lower row. The style encoder improves

Fig. 13: Colorization results with different inference mode. **Style enhance** modes transfer low-level features globally, while **background enhance** modes mainly enhances the backgrounds. We highlight the differences of stroke details inside foregrounds between **style** and **background** modes of rows (2) and (3) in column f. High-resolution images and failure cases are available in the supplementary materials.

the foreground colorization quality in (C), but the results still exhibit additional objects in the upper row and strange colors in the lower row. The full framework with background encoder added in (D) further addresses the issues in (C) and synthesizes visually pleasant and artifacts-free results. The results of the model with all components jointly trained in (E), however, suffer from severe background artifacts, which indicates that different modules are influencing each other and multi-stage training is necessary.

Training strategy for style encoder. Aside from diminishing background artifacts and enabling different inference modes for various use cases, the multi-stage training approach is crucial for disentangling style modulation from colorization optimization. To underscore the significance of the multi-stage strategy, we conducted an ablation study. A baseline model was trained for an equivalent GPU time as our full-stage training, but with the style encoder jointly optimized alongside the sketch encoder and the denoising U-Net in the first stage, rather than separately in the final stage. A comparison is presented in Figure 12. Such joint optimization is problematic because style modulation can hinder the optimization of embedding transfer, as style features stem from low-level visual information and inadvertently encode identity and color semantics. This increased susceptibility typically results in the jointly-trained model performing worse than the model trained via our multi-stage approach during inference, since low-level features are more likely to overfit to training data.

Inference modes. Figure 13 illustrates the distinctions between inference modes. The *vanilla* mode shown in column c relies solely on local image embeddings from the CLIP image encoder as colorization references. Even if the reference images are high-quality high-resolution, the colorization results still exhibit blur textures and simple composition. The *Style enhancement* mode shown in column d synthesizes rich textures and fine details especially for the foreground regions. The *Background enhancement* mode shown in column e loyally reconstructs the color, tone and textures from the reference images in the background regions, and the *Full enhancement* mode shown in column g combines the strength of Style enhancement mode and Background enhancement mode, leading to high-quality artifacts free results with both detailed characters with fine textures and loyally reconstructed background regions.

E. User Study.

To further reveal the subjective evaluation of the proposed method and existing methods by real persons, we demonstrate a user study with 30 participants from Anime lovers communities invited to select the best results with two criteria: the overall colorization quality and the preservation of the geometric structure of the sketches. 26 image sets are prepared, and each participant is shown 15 image sets for evaluation. We present to participants the colorization results of the proposed method and those generated by six existing methods for

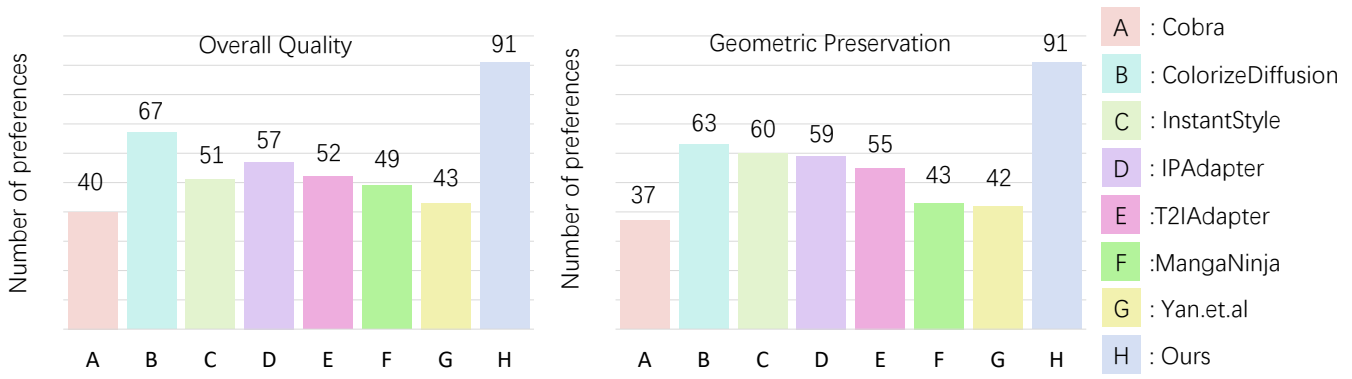

Fig. 14: Results of user study. Our method is preferred across all shown methods in overall quality and geometric preservation.

each image set. We present the results of the user study in Figure 14, with the results showing that our proposed method has received the most numbers of preferences across all the methods illustrated. For further validation of the comparison, the Kruskal-Wallis test is employed as a statistical method. The results demonstrate that our proposed method outperforms all previous methods significantly in terms of user preference with a significance level of $p < 0.05$. All the images shown in the user study are included in the supplementary materials.

V. Conclusion

This paper presents an image-guided sketch colorization framework designed to achieve high-quality and artifacts-free results with arbitrary inputs. To address the limitations of existing methods, we carefully analyzed the sketch colorization workflow in professional animation studios and the intermedium to transfer information from reference to sketches. Based on our analysis, we introduce a multi-stage framework with split cross-attention to separately process the foreground regions and background regions, and employ a background encoder for background feature transfer and a style encoder for style enhancement. This architecture enables zero-shot colorization with state-of-the-art quality and references in not only animation but also various other styles.

However, this work still faces several limitations. First, Our method still faces difficulties in disentangling the sketch-guided foregrounds, such as high-precision character colorization. It is challenging to accurately apply corresponding color references to small regions such as facial characters and detailed clothes, which are discussed in details in the supplementary materials. Also, this work does not consider the inter-frame stabilization and consistency of videos, which should be discussed in future works.

References

[1] D. Yan, X. Wang, Z. Li, S. Saito, Y. Iwasawa, Y. Matsuo, and J. Guo, "Image referenced sketch colorization based on animation creation workflow," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 23 391–23 400.

[2] L. Zhang, Y. Ji, X. Lin, and C. Liu, "Style transfer for anime sketches with enhanced residual u-net and auxiliary classifier gan," in *2017 4th IAPR Asian conference on pattern recognition (ACPR)*. IEEE, 2017, pp. 506–511.

[3] L. Zhang, C. Li, T. Wong, Y. Ji, and C. Liu, "Two-stage sketch colorization," *ACM Trans. Graph.*, vol. 37, no. 6, p. 261, 2018.

[4] C. Zou, H. Mo, C. Gao, R. Du, and H. Fu, "Language-based colorization of scene sketches," *ACM Trans. Graph.*, vol. 38, no. 6, 2019.

[5] D. Yan, R. Ito, R. Moriai, and S. Saito, "Two-step training: Adjustable sketch colourization via reference image and text tag," *Computer Graphics Forum*, 2023.

[6] D. Yan, L. Yuan, E. Wu, Y. Nishioka, I. Fujishiro, and S. Saito, "Colorizediffusion: Improving reference-based sketch colorization with latent diffusion model," in *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 5092–5102.

[7] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *CoRR*, vol. abs/2308.06721, 2023.

[8] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," *CoRR*, vol. abs/2302.08453, 2023.

[9] Z. Liu, K. L. Cheng, X. Chen, J. Xiao, H. Ouyang, K. Zhu, Y. Liu, Y. Shen, Q. Chen, and P. Luo, "Manganinja: Line art colorization with precise reference following," *arXiv preprint arXiv:2501.08332*, 2025.

[10] J. Zhuang, L. Li, X. Ju, Z. Zhang, C. Yuan, and Y. Shan, "Cobra: Efficient line art colorization with broader references," 2025. [Online]. Available: https://arxiv.org/abs/2504.12240

[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *ICLR*. OpenReview.net, 2022.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.

[13] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *ICLR*. OpenReview.net, 2021.

[14] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *ICML*, vol. 37. JMLR.org, 2015, pp. 2256–2265.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014, pp. 2672–2680.

[16] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*. IEEE/CVF, 2019, pp. 4401–4410.

[17] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *CVPR*. IEEE/CVF, 2020, pp. 8107–8116.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, vol. 9351. Springer, 2015, pp. 234–241.

[19] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *ICCV*. IEEE, 2023, pp. 4172–4182.

[20] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, and Z. Li, "Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," 2023.

[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*. IEEE/CVF, 2022, pp. 10 674–10 685.

[22] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: improving latent diffusion models for high-resolution image synthesis," *CoRR*, vol. abs/2307.01952, 2023.

[23] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *ICLR*. OpenReview.net, 2021.

[24] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps," in *NeurIPS*, 2022.
[25] ——, "Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models," *CoRR*, vol. abs/2211.01095, 2022.
[26] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *NeurIPS*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022.
[27] P. Dhariwal and A. Q. Nichol, "Diffusion models beat gans on image synthesis," in *NeurIPS*, 2021, pp. 8780–8794.
[28] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *CoRR*, vol. abs/2207.12598, 2022.
[29] G. Kwon and J. C. Ye, "Diffusion-based image translation using disentangled style and content representation," in *ICLR*. OpenReview.net, 2023.
[30] H. Wang, Q. Wang, X. Bai, Z. Qin, and A. Chen, "Instantstyle: Free lunch towards style-preserving in text-to-image generation," *arXiv preprint arXiv:2404.02733*, 2024.
[31] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu, "Inversion-based style transfer with diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 146–10 156.
[32] Y. Cao, X. Meng, P. Y. Mok, T.-Y. Lee, X. Liu, and P. Li, "Animediffusion: Anime diffusion colorization," *TVCG*, pp. 1–14, 2024.
[33] D. Yan, L. Yuan, Y. Nishioka, I. Fujishiro, and S. Saito, "Colorizediffusion: Adjustable sketch colorization with reference image and text," 2024.
[34] B. Zhang, Y. Duan, J. Lan, Y. Hong, H. Zhu, W. Wang, and L. Niu, "Controlcom: Controllable image composition using diffusion model," *arXiv preprint arXiv:2308.10040*, 2023.
[35] K. Kim, S. Park, J. Lee, and J. Choo, "Reference-based image composition with sketch via structure-aware diffusion model," *arXiv preprint arXiv:2304.09748*, 2023.
[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
[37] L. Hu, X. Gao, P. Zhang, K. Sun, B. Zhang, and L. Bo, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," *arXiv preprint arXiv:2311.17117*, 2023.
[38] Z. Huang, M. Zhang, and J. Liao, "LVCD: reference-based lineart video colorization with diffusion models," *ACM Trans. Graph.*, vol. 43, no. 6, pp. 177:1–177:11, 2024.
[39] J. Xing, H. Liu, M. Xia, Y. Zhang, X. Wang, Y. Shan, and T. Wong, "Tooncrafter: Generative cartoon interpolation," *ACM Trans. Graph.*, vol. 43, no. 6, pp. 245:1–245:11, 2024.
[40] Z. Li, Z. Geng, Z. Kang, W. Chen, and Y. Yang, "Eliminating gradient conflict in reference-based line-art colorization," in *ECCV*. Springer, 2022, pp. 579–596.
[41] X. Wang and J. Yu, "Learning to cartoonize using white-box cartoon representations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8090–8099.
[42] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "Cartoongan: Generative adversarial networks for photo cartoonization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9465–9474.
[43] L. Zhang, X. Wang, Q. Fan, Y. Ji, and C. Liu, "Generating manga from illustrations via mimicking manga creation workflow," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5642–5651.
[44] D. Sýkora, J. Dingliana, and S. Collins, "Lazybrush: Flexible painting tool for hand-drawn cartoons," *Comput. Graph. Forum*, vol. 28, no. 2, pp. 599–608, 2009.
[45] H. Kim, H. Y. Jhoo, E. Park, and S. Yoo, "Tag2pix: Line art colorization using text tag with secat and changing loss," in *ICCV*. IEEE/CVF, 2019, pp. 9055–9064.
[46] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023, pp. 3836–3847.
[47] L. Zhang, "Style2paints v5," 2023, accessed: DATE 2023-06-25.
[48] Kohya-ss, "Hugging face/controlnet-lllite," https://huggingface.co/kohya-ss/controlnet-lllite, 2024, accessed: DATE 2024-01-02.
[49] L. Zhang, "Controlnet-v1-1-nightly," https://github.com/lllyasviel/ControlNet-v1-1-nightly, 2024, accessed: DATE 2024-01-02.
[50] Y. Meng, H. Ouyang, H. Wang, Q. Wang, W. Wang, K. L. Cheng, Z. Liu, Y. Shen, and H. Qu, "Anidoc: Animation creation made easier," *arXiv preprint arXiv:2412.14173*, 2024.
[51] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, vol. 139. PMLR, 2021, pp. 8748–8763.
[52] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *CVPR*, 2023, pp. 2818–2829.
[53] C. Schuhmann, R. Beaumont, R. Vencu, C. W. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. R. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. 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.
[54] SkyTNT, infoengine1337, and not lain, "anime-segmentation," https://github.com/SkyTNT/anime-segmentation, 2022.
[55] X. Qin, H. Dai, X. Hu, D.-P. Fan, L. Shao, and L. V. Gool, "Highly accurate dichotomous image segmentation," in *ECCV*, 2022.
[56] D. Yan, X. Wang, Z. Li, S. Saito, Y. Iwasawa, Y. Matsuo, and J. Guo, "Enhancing reference-based sketch colorization via separating reference representations," *arXiv preprint arXiv:2508.17620*, 2025.
[57] D. Yan, X. Wang, Z. Li, S. Saito, Y. Iwasawa, Y. Matsuo, and J. Guo, "Image Referenced Sketch Colorization Based on Animation Creation Workflow," *arXiv e-prints*, 2025.
[58] X. Huang and S. J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *ICCV*. IEEE/CVF, 2017, pp. 1510–1519.
[59] D. community, G. Branwen, and Anonymous, "Danbooru2021: A large-scale crowdsourced and tagged anime illustration dataset," https://gwern.net/danbooru2021, 2022, accessed: DATE 2022-01-21.
[60] L. Zhang, "Sketchkeras," https://github.com/lllyasviel/sketchKeras, 2017.
[61] X. Xiang, D. Liu, X. Yang, Y. Zhu, X. Shen, and J. P. Allebach, "Adversarial open domain adaptation for sketch-to-photo synthesis," in *WACV*. IEEE/CVF, 2022, pp. 944–954.
[62] Microsoft, "Deepspeed," https://github.com/microsoft/DeepSpeed, 2024.
[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
[64] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*. OpenReview.net, 2019.
[65] discus0434, "aesthetic-predictor-v2-5," https://github.com/discus0434/aesthetic-predictor-v2-5, 2024, accessed: DATE 2024-12-06.
[66] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *NeurIPS*, 2017, pp. 6626–6637.

Dingkun Yan received his B.Eng. in The school of Computer Science and Technology from East China Normal University in 2019, and Ph.D degree in Computer Science from the Tokyo Institute of Technology in 2025. His research interests include reference-based sketch colorization, non-photorealistic rendering for anime/manga, and diffusion-based image synthesis.

Xinrui Wang is a PhD candidate from the Department of Technology Management of Innovation, Graduate School of Engineering, The University of Tokyo. He received his B.Eng degree in engineering from University of Science and Technology (USTC) in 2016 and M.Eng degree in engineering from The University of Tokyo in 2018. His research interests include image synthesis, image editing, image and video stylization, generative models, and ACG related machine learning applications. For more details, please refer to https://github.com/SystemErrorWang.

Zhuoru Li received his PhD degree in chemistry from Xiamen Universty, China. Currently, he is working at Project HAT. His current research interests include computer graphics, non-photorealistic rendering and artificial intelligence for science..

D r. Yu Jinze is a Senior Applied Scientist at Amazon Japan, specializing in multimodal large language models and generative AI technologies for images, speech, and music. He is also a visiting researcher of Department of Communications and Computer Engineering at Waseda University. He holds a Bachelor's degree in Mathematics from Tsinghua University, a Master's in Computer Engineering from École Polytechnique Paris, and a Ph.D. in Electronics and Information Engineering from the University of Tokyo. He has published papers in top-tier conferences and journals including CVPR and IEEE TVCG, and holds more than authorized patents across China, the United States, and Japan as the first inventor. His research interests also cover AI perception and decision-making systems for complex industrial environments, with particular emphasis on multimodal sensor fusion and intelligent control algorithms for robotics applications.

Suguru Saito received his B.Eng. degree in computer science in 1994 and Ph.D. degree in 1999, both from the Tokyo Institute of Technology. He is currently an Associate Professor with the School of Computing, Institute of Science Tokyo (formed by the 2024 merger; formerly Tokyo Institute of Technology), Tokyo, Japan. His research interests include computer graphics, non-photorealistic rendering, foveated/peripheral-vision-aware rendering, image processing, and perceptual modeling for visualization and animation. He has led multiple KAKENHI projects on peripheral-vision-aware graphics and anime production data workflows, and has published in venues spanning graphics and imaging.

Yusuke Iwasawa received his bachelor degree of engineering and master degree of engineering from Sophia University in 2012 and 2014, and recieved PhD from Graduate School of Engineering, The University of Tokyo at 2017. After working as special researcher, special teaching assistant and special lecture, he started to work as an associate professor at Department of TMI, Graduate School of Engineering, The University of Tokyo from April, 2022. His research fields include wearable sensing, deep learning and applications of machine learning.

Yutaka Matsuo is a professor at Graduate School of Engineering, the University of Tokyo. He received his BS, MS, and Ph.D. degrees from the University of Tokyo in 1997, 1999, and 2002. After working at National Institute of Advanced Industrial Science and Technology (AIST) and Stanford University, he joined the faculty of University of Tokyo in 2007. At Japan Society for Artificial Intelligence (JSAI), he has served as Editor-in-chief and the chair of the ELSI committee, and has been a board member since 2020. He is the president of Japan Deep Learning Association (JDLA), and a member of the board of directors at SoftBank Group Corp. He is working on artificial intelligence, especially on deep learning and web mining.

Jiaxian Guo is a senior research scientist at the Google Research. He completed his bachelor's degree at Shanghai Jiao Tong University in 2018 and his PhD from the University of Sydney in 2022. He was a Postdoc in the University of Tokyo in 2023. He has published several papers and serves as a reviewer for prestigious conferences such as NeurIPS, ICML, CVPR, and ICLR. His research interests lie in causality-inspired generative models and decision-making. His several works have been recognized by several medias. Jiaxian's academic excellence was highlighted by his nomination for the best thesis candidate at Shanghai Jiao Tong University and receiving the PhD Completion Award from the University of Sydney.