One-shot Embroidery Customization via Contrastive LoRA Modulation

JUN MA, Zhejiang Sci-Tech University and Style3D Research, China
QIAN HE*, State Key Lab of CAD&CG, Zhejiang University and Style3D Research, China
GAOFENG HE, Style3D Research, China
HUANG CHEN, Style3D Research, China
CHEN LIU, State Key Lab of CAD&CG, Zhejiang University and Style3D Research, China
XIAOGANG JIN, State Key Lab of CAD&CG, Zhejiang University, China
HUAMIN WANG, Style3D Research, China

Fig. 1. Given a reference embroidery image (Columns 1 and 4), our method is capable of generating novel embroidery images (Columns 3 and 5) based on either image inputs (Column 2) or textual descriptions (Column 5). These outputs can be seamlessly integrated with ACE++ [Mao et al. 2025] to produce decorative images for virtual display applications (Column 6). Furthermore, our approach demonstrates strong generalization across a range of visual attribute transfer tasks, including artistic style transfer (Row 1, last column), sketch colorization (Row 2, last column), and appearance transfer (Row 3, last column). Pink flower design (Row 3, Column 2) © Vecteezy.

*Project lead and corresponding author.

Authors' addresses: Jun Ma, majun88818@163.com, Zhejiang Sci-Tech University and Style3D Research, Hangzhou, China; Qian He, heqianhailie@gmail.com, State Key Lab of CAD&CG, Zhejiang University and Style3D Research, Hangzhou, China; Gaofeng He, hegaofeng@linctex.com, Style3D Research, Hangzhou, China; Huang Chen, chenhuang@linctex.com, Style3D Research, Hangzhou, China; Chen Liu, eric.liu@linctex.com, State Key Lab of CAD&CG, Zhejiang University and Style3D Research, Hangzhou, China; Xiaogang Jin, jin@cad.zju.edu.cn, State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China; Huamin Wang, wanghmin@gmail.com, Style3D Research, Hangzhou, China.

ACM 0730-0301/2025/12-ART271
https://doi.org/10.1145/3763290

Diffusion models have significantly advanced image manipulation techniques, and their ability to generate photorealistic images is beginning to transform retail workflows, particularly in presale visualization. Beyond artistic style transfer, the capability to perform fine-grained visual feature transfer is becoming increasingly important. Embroidery is a textile art form characterized by intricate interplay of diverse stitch patterns and material properties, which poses unique challenges for existing style transfer methods. To explore the customization for such fine-grained features, we propose a novel contrastive learning framework that disentangles fine-grained style and content features with a single reference image, building on the classic concept of image analogy. We first construct an image pair to define the target style, and then adopt a similarity metric based on the decoupled representations of pretrained diffusion models for style-content separation. Subsequently, we propose a two-stage contrastive LoRA modulation technique to capture fine-grained style features. In the first stage, we iteratively update the whole LoRA and the selected style blocks to initially separate style from content. In the second stage, we design a contrastive learning strategy to further decouple style and content through self-knowledge distillation. Finally, we build an inference pipeline to handle image or text inputs with only the style blocks. To evaluate our method on fine-grained

style transfer, we build a benchmark for embroidery customization. Our approach surpasses prior methods on this task and further demonstrates strong generalization to three additional domains: artistic style transfer, sketch colorization, and appearance transfer. Our project is available at: https://style3d.github.io/embroidery_customization.

CCS Concepts: • **Computing methodologies** → **Image manipulation**; **Machine learning approaches**.

Additional Key Words and Phrases: Embroidery customization, one-shot, low-rank adaptation, contrastive learning

ACM Reference Format:
Jun Ma, Qian He, Gaofeng He, Huang Chen, Chen Liu, Xiaogang Jin, and Huamin Wang. 2025. One-shot Embroidery Customization via Contrastive LoRA Modulation. *ACM Trans. Graph.* 44, 6, Article 271 (December 2025), 15 pages. https://doi.org/10.1145/3763290

1 INTRODUCTION

Visual attribute transfer [Efros and Freeman 2001; Hertzmann et al. 2001; Liao et al. 2017] represents a fundamental challenge in image manipulation, involving the separation and recombination of style and content [Tenenbaum and Freeman 1996], and is revitalized by recent advances in diffusion-based generative models [Podell et al. 2023; Rombach et al. 2022], particularly in the domain of artistic style transfer [Wang et al. 2024; Zhang et al. 2023b; Zhou et al. 2025]. Beyond virtual display, diffusion models now generate high-quality, photorealistic images from customized instructions, often outperforming traditional 3D modeling and rendering. This capability is beginning to transform and even revolutionize retail workflows under the "sell it before you make it" paradigm [Lin et al. 2025], offering a novel approach to inventory challenges. However, controllability over fine-grained structural elements such as embroidery or real-world textiles remains a key challenge.

Embroidery is an intricate textile art characterized by the structured arrangement of diverse yarns and materials, as shown in Fig. 1. Customization of embroidery styles poses unique challenges for existing methods in visual style transfer. To begin with, relying solely on pretrained models often fails to generalize to unseen embroideries [Chung et al. 2024; Wang et al. 2024], while fine-tuning on large-scale datasets [Qi et al. 2024; Xing et al. 2024] also proves ineffective due to data scarcity and complex intra-class variation in embroidery patterns. Furthermore, general style transfer often treats color as a key component of style [Frenkel et al. 2025], whereas embroidery style focuses on high-frequency structural textures, largely independent of color, causing existing network block selection methods to struggle with separating embroidery style from its graphic design content. Additionally, other effective constraints for general style disentanglement can still have difficulty in capturing complex embroidery styles [Jones et al. 2024], as shown in Fig. 2.

To address these challenges, we propose a novel framework for fine-grained style customization, with embroidery as a representative case. Our main idea is to employ contrastive learning with a single reference image to achieve style–content disentanglement, which comprises three aspects: Firstly, we revisit the classic concept from image analogy [Efros and Freeman 2001] by constructing a single image pair to define a style, to reduce ambiguity and avoid inconsistency among different style images. Secondly, we adopt a metric to measure feature similarity within an image pair and cross

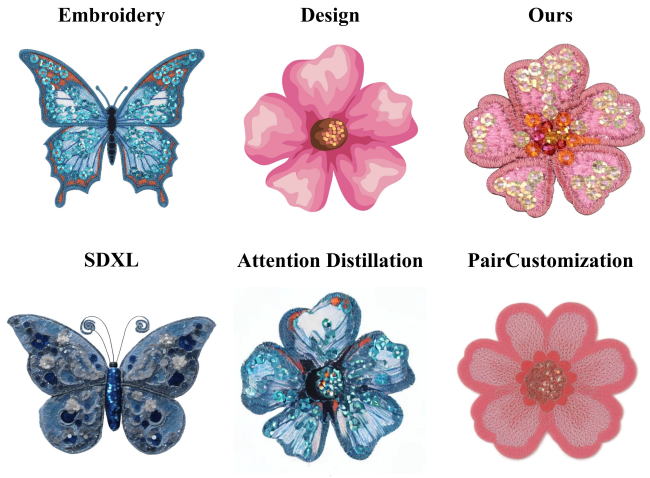


Fig. 2. Challenges in embroidery customization. Given a reference embroidery, SDXL [Podell et al. 2023] generates dissimilar style with merely text input "blue butterfly embroidery patch with sequins". Attention Distillation [Zhou et al. 2025] and PairCustomization [Jones et al. 2024] both fail to capture sequin structure and yarns while maintainig the pink flower design, comparing to Ours. Pink flower design (Row 1, Column 2) © Vecteezy.

image pairs, leveraging the decoupled representations [Liao et al. 2017] of pretrained diffusion models. Finally, we design a contrastive LoRA [Ryu 2022] modulation technique named EmoLoRA, to first capture style features in selected blocks, and then further decouple style from content with self-knowledge distillation.

Specifically, our framework comprises the following steps: pairwise data construction, analysis on SDXL, contrastive LoRA learning and model inference. For a reference embroidery image, we leverage the prior knowledge in SD3 [Esser et al. 2024] combined with ControlNets [Zhang et al. 2023c], to generate its graphic design image. Then we analyze SDXL by computing the cosine similarity between the constructed image pair in an inversion-reconstruction pipeline [Garibi et al. 2024], using self-attention output features in each transformer block. Blocks with low feature similarity demonstrate closer correlation to the specified style, which is the main distinction between the image pair. After data preparation and model analysis, we design a two-stage contrastive LoRA learning strategy, to alleviate the overfitting and "content leakage" issue of the standard LoRA. In the first stage, we iteratively update the whole LoRA and the selected style blocks with supervision on the generated graphic design and the reference embroidery image, respectively. This modulation mechanism constrains the style into the selected blocks while leaves major content features to the other blocks. In the second stage, we use the trained LoRA to generate complementary data and apply contrastive learning on the noised latent feature space [Dalva and Yanardag 2024], to further decouple style and content in the style blocks. Finally, we build a model inference pipeline to handle image or text inputs with only the style blocks.

Our main contributions are summarized as follows:

- We propose a novel contrastive learning framework for fine-grained style customization, by constructing an image pair to

define a style and designing a contrastive LoRA modulation technique to decouple style and content.
- We introduce a new task, one-shot embroidery customization, which poses unique challenges with intricate structural features, and conduct analysis to verify its potential in transforming real-world embroidery workflows.
- We outperform existing approaches in separating embroidery style from design content and exhibit strong generalization to three additional domains: artwork-photo, color-sketch, and appearance-structure.

2 RELATED WORK

Diffusion-based Image Synthesis. Image synthesis has achieved tremendous progress with the rise of diffusion-based generative models [Dhariwal and Nichol 2021; Ho et al. 2020; Peebles and Xie 2023; Rombach et al. 2022; Sohl-Dickstein et al. 2015]. Harnessing the generative power of pre-trained text-to-image models, various applications in personalization / customization [Kumari et al. 2023; Tang et al. 2024; Tewel et al. 2023; Zhang et al. 2023a] are developed. Given a small image set in a new concept, Dreambooth [Ruiz et al. 2023] adapts the model via finetuning, while TI [Gal et al. 2022] finds the embeddings in the textual feature space. LoRA is a PEFT method [Houlsby et al. 2019] to adapt large language [Hu et al. 2021] or vision [Ryu 2022] models to downstream tasks. Recent works [Mou et al. 2024; Ye et al. 2023; Zhang et al. 2023c] propose plug-and-play adapters that enable controllable image generation by modulating the generative process without retraining the base model. To handle image editing and translating [Brooks et al. 2023; Kawar et al. 2023; Nichol et al. 2021; Valevski et al. 2023], SDEdit [Meng et al. 2021] first adds noise to the input image and then denoises it through the SDE prior, while [Hertz et al. 2022; Tumanyan et al. 2023] adopt an inversion [Mokady et al. 2023; Song et al. 2020] pipeline and attention feature manipulation [Liu et al. 2024]. Due to the limited expressiveness of text for fine-grained spacial features, we mainly explore LoRA-based methods to leverage the representation capacity of pretrained diffusion models, and propose to further decouple these features for style customization during finetuning.

Visual Attribute Transfer. Visual attribute transfer [Efros and Freeman 2001; Hertzmann et al. 2001] aims to transform an image to adopt the style of another, encompassing elements such as color, texture, local structures, and artistic style. Neural style transfer approaches have evolved from early CNN-based frameworks [Gatys et al. 2016; Huang and Belongie 2017; Johnson et al. 2016; Li et al. 2017; Park and Lee 2019; Zhang et al. 2022], to adversarial learning with GANs [Goodfellow et al. 2014; Isola et al. 2017; Karras 2019; Karras et al. 2020; Park et al. 2020; Zhu et al. 2017], and more recently to transformer architectures leveraging self-attention for global context [Wu et al. 2021]. The rapid development of text-to-image diffusion models has also sparked their adaptation to style transfer tasks, including exploring the textual feature space [Li et al. 2025; Qi et al. 2024; Yang et al. 2023; Zhang et al. 2023b], manipulating the attention features [Chung et al. 2024; Deng et al. 2023; Hertz et al. 2024], leveraging plug-and-play adapters [Wang et al. 2024], finetuning LoRAs [Frenkel et al. 2025; Jones et al. 2024; Shah et al. 2025], and formulating the problem using stochastic optimal control [Rout et al. 2024] with an existing style descriptor [Somepalli et al. 2024]. However, these methods fall short when handling intricate structural styles such as embroidery, where color serves as content rather than style, and high-frequency structural textures, commonly neglected in artistic style transfer, play a central role in characterizing style. A detailed discussion is provided in Sec. 4.2. Sketch colorization [Li et al. 2022; Yan et al. 2025; Zhang et al. 2021] and appearance transfer [Alaluf et al. 2024; Kwon and Ye 2022; Tumanyan et al. 2022] are slightly different problems than artistic style transfer, as they have different definition of style and content, and potentially involve semantic correspondence between reference and target images. Attention Distillation [Zhou et al. 2025] transfers style or appearance using distillation loss on pretrained attention features, yet still has difficulty in separating structural embroidery styles from color content, as presented in Fig. 2. Furthermore, diffusion-based image analogy frameworks propose a more generic approach for visual attribute transfer. DIA [Šubrtová et al. 2023] focuses on high-level semantics and represents $A : A'$ in CLIP embedding space, while Analogist [Gu et al. 2024] uses GPT-4V [Achiam et al. 2023] to reason the analogy $A : A' :: B : B'$ and relies on textual descriptions to capture style, both are limited in capturing fine-grained styles like embroidery. In this work, we propose a novel framework to capture intricate structural styles at a fine-grained level, thereby addressing challenges more closely aligned with real-world applications.

Embroidery Synthesis. Embroidery is a decorative fabric art form [Nichols 2012; Pile 2018] that can take on various styles, each exhibiting distinct visual characteristics due to the use of different yarns and stitches, as well as materials such as pearls, beads, and sequins (Fig. 1 and 5). For example, chenille embroidery creates a fuzzy, textured surface with looped yarn (Row 2, Column 4 in Fig. 1), while sequin embroidery incorporates reflective discs secured by stitches for added sparkle (Row 3, Column 1 in Fig. 1). Contemporary embroidery design heavily relies on specialized CAD software (e.g., Wilcom EmbroideryStudio[1]) that translates digital artwork into machine-readable stitch instructions, but the process still involves extensive manual operation and often lacks fully satisfactory visualizations. Automated embroidery synthesis has continued to attract interest as a specialized but intriguing topic in computer graphics and digital fabrication. Early approaches model flat embroidery in 3D by incorporating three fundamental stitch types—long-short, satin, and stem/edge stitches—into geometric representations [Chen et al. 2012; Cui et al. 2017]. Subsequent work focuses on simulating random-needle embroidery using vector fields or stitch primitives, followed by multilayer rendering techniques to enhance visual fidelity [Ma and Sun 2022; Yang and Sun 2018; Yang et al. 2012]. To improve rendering realism, intrinsic image decomposition is introduced to preserve the illumination of input photographs during the synthesis process [Shen et al. 2017]. The manual specification of stitch types is further alleviated by segmenting input images into subregions, assigning appropriate stitch categories to each segment, and synthesizing embroidery textures via UV mapping [Guan et al. 2021]. More recent approaches leverage deep neural networks and adversarial

[1] https://wilcom.com/

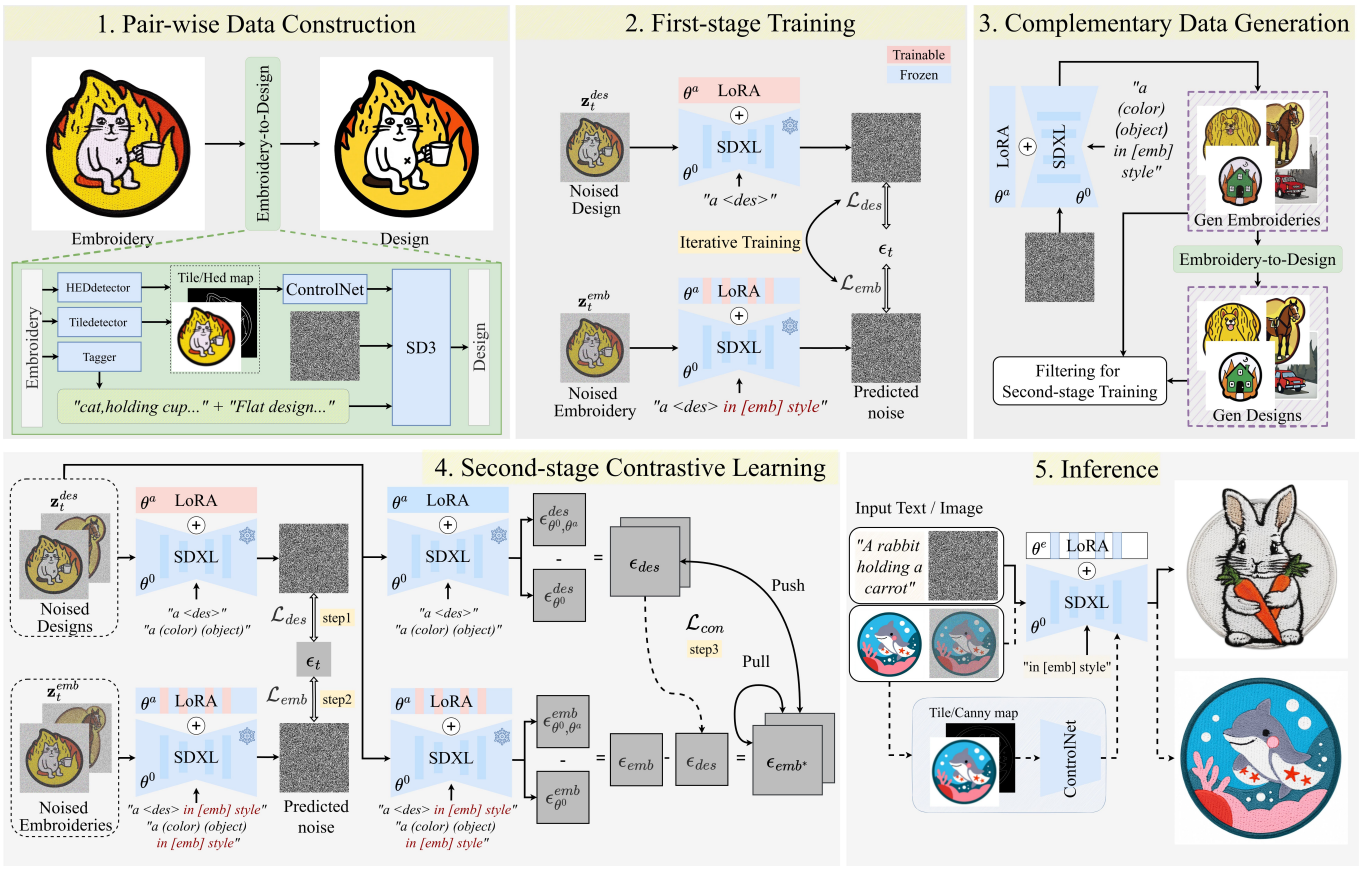


Fig. 3. An overview of our framework. (1) Pair-wise data construction: We build a pipeline to process the reference embroidery into its graphic design; (2) First-stage training: Given the embroidery-design pair, we iteratively train a LoRA θ^a to initially decouple style and content; (3) Complementary data generation: We generate more embroidery-design pairs with the trained LoRA from the first stage; (4) Second-stage contrastive learning: We conduct contrastive learning to further decouple style and content in embroidery LoRA blocks θ^e using the reference and a generated emb-des pair; (5) Model inference: We use the trained embroidery LoRA θ^e to conduct text/image-based customization.

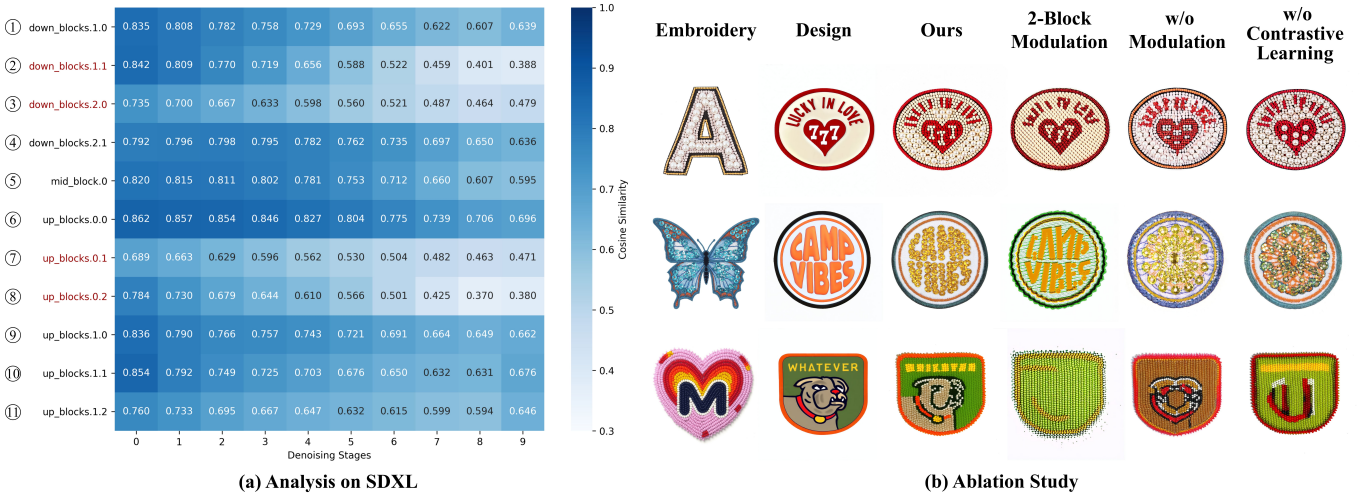
learning to directly generate embroidery-like imagery [Yang et al. 2022; Ye et al. 2021]. Empowered by supervised deep learning, stitch types in segmented subregions can be explicitly classified using annotated datasets [Hu et al. 2024]. Beyond visual realism, recent efforts have also explored the generation of machine-fabricable embroidery patterns. For example, Liu et al. [2023] propose a method that uses user-defined directional cues and vector field analysis to derive continuous streamlines suitable for machine stitching. In contrast to prior methods, our approach takes as input a reference embroidery image and a natural language prompt, and generates a customized embroidery image. It supports a wide variety of stitches and materials without relying on explicit stitch-type labeling or manual annotations.

3 METHOD

In this section, we introduce our contrastive learning framework for fine-grained style customization, using embroidery as a representative case. Given a reference embroidery image I, our objective is to generate embroidery images that replicate the same style, encompassing stitches, yarns, accessories, and other prominent structural features. To enable contrastive learning with a single reference image, we first introduce a pair-wise data construction module for style definition in Sec. 3.1 and a similarity metric for identifying style features in Sec. 3.2. Building on this foundation, we then present our two-stage contrastive LoRA learning strategy for style-content disentanglement in Sec. 3.3. Finally, we describe our model inference in Sec. 3.4. An overview of our framework is in Fig. 3.

3.1 Pair-wise Data Construction

Given a single reference embroidery image, our goal is to disentangle style from content to provide supervision for contrastive learning. While embroidery style is abstract and difficult to capture explicitly, design content is comparatively easier to represent. To this end, we construct a data pipeline that generates a corresponding graphic design image, thereby defining style through a data pair in the spirit of image analogy [Efros and Freeman 2001]. In this embroidery-to-design module, we adopt the text-to-image



(a) Analysis on SDXL **(b) Ablation Study**

Fig. 4. Base model analysis and ablation study. (a) Average cosine similarity of self-attention outputs between all paired embroidery and design images in our reference set. A cosine similarity of 1 indicates that the block produces nearly identical features for an embroidery-design pair, implying that it primarily represents non-embroidery content. (b) Ablation study on three reference embroidery examples.

pipeline of SD3 [Esser et al. 2024] with ControlNet [Zhang et al. 2023c]. Specifically, we utilize the HEDdetector [Xie and Tu 2015] to detect rough edges of the embroidery image, and then send the edge image to ControlNet-Canny. In this way, the design structure is well preserved, while the embroidery stitches are effectively removed. To preserve color fidelity and enhance generation quality, we send the blurred embroidery image into a ControlNet-Tile branch. Moreover, we use WD14 [SmilingWolf 2023] to generate captions as the prompt with "*flat design, vector graphic design, digital design, cartoon design, clean lines, uniform color blocks, smooth surface, high quality*" and thus achieve better preservation of the design content. Empirically, we find that using the pipeline with SD3 yields better results than SDXL, which is probably due to differences in their pretraining datasets. While our embroidery-to-design module may not generalize directly to other styles, the underlying concept of pair-wise data construction can be adapted through alternative means.

3.2 Analysis on SDXL

Inspired by B-LoRA [Frenkel et al. 2025], we leverage the decoupled representations of pretrained diffusion models and employ different LoRA blocks to separate style and content. Unlike B-LoRA, however, embroidery style cannot be captured by a single block entangled with color information, as it is largely independent of color, nor can a single content block fully reconstruct the detailed design. Building on prior findings [Liu et al. 2024; Tumanyan et al. 2023] that self-attention features in the UNet of diffusion models encode spatial structure, including high-frequency details, we introduce a similarity metric to guide the selection of network blocks most suitable for capturing a specified style. We adopt SDXL as our base model for its higher resolution, improved visual fidelity, and more naturally decoupled attention features compared to SD3. In this work, we employ ReNoise [Garibi et al. 2024], an inversion technique that achieves higher reconstruction quality than DDIM [Song et al. 2020], and leverage the output features of each self-attention layer in the image reconstruction process to compare differences between each embroidery-design pair. Let $\mathbf{F}_i$ denote the input feature to a self-attention layer, and $f_i^q(\cdot)$, $f_i^k(\cdot)$, $f_i^v(\cdot)$ and $f_i^o(\cdot)$ be the projection layers for query, key, value and output, respectively. We have $\mathbf{Q}_i = f_i^q(\mathbf{F}_i)$, $\mathbf{K}_i = f_i^k(\mathbf{F}_i)$, $\mathbf{V}_i = f_i^v(\mathbf{F}_i)$, and d_k as the dimensionality of query and key. The output self-attention feature is:

$$\mathbf{F}_i^o = f_i^o(Softmax(\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{d_k}})\mathbf{V}_i). \quad (1)$$

For each block, we compute the average cosine similarity of all $\mathbf{F}_i^o$ between an embroidery-design pair. Since the whole generation process consists of 50 steps, we divide them into 10 sections and compute the average. Empirically, we find the similarity matrix for the collected reference embroidery set share the same relative scale and thus adopt their average as shown in Fig. 4 (a), where a cosine similarity of 1 indicates that the block produces nearly identical features for an embroidery-design pair, implying that it primarily represents non-embroidery content. We notice that *down_blocks 1.1, 2.0* and *up_blocks 0.1, 0.2* show more correlation to embroidery features as they differ more between embroidery and design. The difference is more significant in later stages of the denoising process, mainly because embroidery style is more related to delicate low-level image features than rough high-level semantic features.

3.3 Contrastive LoRA Learning

With the constructed embroidery-design pairs and network feature analysis, we now present a two-stage contrastive LoRA learning strategy named EmoLoRA, as in Alg. 1, that captures embroidery style from a single reference while alleviating overfitting of standard LoRA [Ryu 2022]. In this work, LoRA is applied to the attention layers of the UNet backbone. During the first stage, we train EmoLoRA

ALGORITHM 1: Two-stage Contrastive EmoLoRA Learning

Input: Constructed embroidery-design image pair (I_{emb}, I_{des}), text pair ("a <des> in [emb] style", "a <des>"), SDXL base model θ^0, EmoLoRA θ^a, selected embroidery blocks θ^e of EmoLoRA, learning rates η_1 and η_2

Output: Trained EmoLoRA θ^a with embroidery blocks θ^e

Stage 1: LoRA Block Modulation

for *each iteration* **do**

Step 1, Update θ^a with I_{des}: $\theta^a \leftarrow \theta^a - \eta_1 \nabla_{\theta^a} \mathcal{L}_{des}$;

Step 2, Update θ^e with I_{emb}: $\theta^e \leftarrow \theta^e - \eta_1 \nabla_{\theta^e} \mathcal{L}_{emb}$;

end

Complementary Data Generation:

Prompt: "a (color) (object) in [emb] style" $\times N$;

Generate embroidery images and select $\lceil N/2 \rceil$;

Obtain design images and select $\lceil N/4 \rceil$;

Stage 2: Contrastive Learning

for *each iteration* **do**

Sample a generated pair $(I^{gen}_{emb}, I^{gen}_{des})$;

Step 1, Update θ^a with I_{des} and I^{gen}_{des}: $\theta^a \leftarrow \theta^a - \eta_1 \nabla_{\theta^a} \mathcal{L}_{des}$;

Step 2, Update θ^e with I_{emb} and I^{gen}_{emb}: $\theta^e \leftarrow \theta^e - \eta_1 \nabla_{\theta^e} \mathcal{L}_{emb}$;

Step 3, Update θ^e with (I_{emb}, I_{des}) and $(I^{gen}_{emb}, I^{gen}_{des})$: $\theta^e \leftarrow \theta^e - \eta_2 \nabla_{\theta^e} \mathcal{L}_{con}$;

end

to roughly decouple embroidery and design through a block modulation mechanism. In the second stage, we use the trained EmoLoRA to generate embroidery images with preset prompts for more supervision signals, and then adopt contrastive learning to further enhance the decoupling of embroidery and design.

LoRA Block Modulation. For our EmoLoRA, we separate the four blocks discussed in Sec. 3.2 to capture embroidery style while using the whole LoRA to recover the design content. Given an embroidery-design pair, we set the prompt for the embroidery image as "*a <des> in [emb] style*", and the prompt for the design as "*a <des>*". During training, we update EmoLoRA weights in a two-step iterative manner. In Step 1, we input "*a <des>*" into all blocks of the SDXL base model θ^0 and EmoLoRA θ^a, and train θ^a with $\mathcal{L}_{des}(\theta^a)$. In Step 2, we input "*a <des> in [emb] style*" into the four embroidery blocks of SDXL and EmoLoRA, and "*a <des>*" into all other blocks, and only update the four embroidery blocks θ^e of EmoLoRA with $\mathcal{L}_{emb}(\theta^e)$. $\mathbf{z}_t^{des}$ and $\mathbf{z}_t^{emb}$ denote the noised image features in latent space [Esser et al. 2021; Kingma 2013], $\mathbf{c}_{des}$ and $\mathbf{c}_{emb}$ are the encoded text features [Radford et al. 2021]. Note that θ^e is a subset of θ^a. We define $\mathcal{L}_{des}(\theta^a)$ and $\mathcal{L}_{emb}(\theta^e)$ as:

$$\mathcal{L}_{des}(\theta^a) = ||\epsilon_t - \epsilon_{\theta^0,\theta^a}(\mathbf{z}_t^{des}, t, \mathbf{c}_{des})||_2^2, \tag{2}$$

$$\mathcal{L}_{emb}(\theta^e) = ||\epsilon_t - \epsilon_{\theta^0,\theta^a}(\mathbf{z}_t^{emb}, t, \mathbf{c}_{emb})||_2^2. \tag{3}$$

After the iterative training process, the embroidery style is encapsulated solely in θ^e and decoupled from the main content, as only θ^e is updated during Step 2. However, θ^e also contains some content information learned from Step 1, which is unable to avoid as the other blocks alone cannot recover the whole content image. Consequently, the generated images may retain color from the reference embroidery and have suboptimal fusion with the new content due to entanglement between the style and its original content, as shown in Fig. 4 (b) **w/o Contrastive Learning**. To further alleviate this problem, we adopt a second stage with contrastive learning.

Complementary Data Generation. Before applying contrastive learning, we generate more data using the trained EmoLoRA from the first stage. We generate new embroidery images with a predefined set of prompts in "*a (color) (object) in [emb] style*", covering various color-object combinations and blending the prior knowledge in SDXL with the learned embroidery style. One example is "*a yellow dog in [emb] style*", while the list of all N prompts is in our supplementary and we set N to 10 in this paper. We then compute the average cosine similarity of each generated image to the reference image using their self-attention output features, as in Sec. 3.2, to measure the style similarity. Since embroidery features are more salient in later generation stages, we only use features from stages 5-9. Then we rank the generated images w.r.t. their average similarity and select the top half $\lceil N/2 \rceil$ for better embroidery quality, and use the embroidery-to-design pipeline in Sec. 3.1 to obtain their corresponding design images. Finally, we want to remove images with content that is too similar to the reference. So we compute and rank the average cosine similarity among the design images similar to embroidery images as before, and choose the most dissimilar half $\lceil N/4 \rceil$ to be our final complementary data.

Contrastive Learning. With the embroidery-design pairs from the initial reference and complementary generation, we now apply contrastive training. The main objective is to pull the embroidery features shared by different image pairs together, and to push away the embroidery features from the content features. Inspired by NoiseCLR [Dalva and Yanardag 2024], we conduct contrastive learning in the noised latent feature space. We obtain the design content features ϵ_{des} by subtracting base model prediction from base model with EmoLoRA prediction, given noised design image features $\mathbf{z}_t^{des}$ and encoded text features $\mathbf{c}_{des}$ at timestep t. Similarly, we can obtain the embroidery image features ϵ_{emb}, while with noised design image features $\mathbf{z}_t^{des}$ but embroidery prompt features $\mathbf{c}_{emb}$. In this way, we are able to separate the learned knowledge in EmoLoRA θ^a that is triggered by "a <des>" or "a <des> in [emb] style". The formulation is as follows:

$$\epsilon_{des} = \epsilon_{\theta^0,\theta^a}(\mathbf{z}_t^{des}, t, \mathbf{c}_{des}) - \epsilon_{\theta^0}(\mathbf{z}_t^{des}, t, \mathbf{c}_{des}), \tag{4}$$

$$\epsilon_{emb} = \epsilon_{\theta^0,\theta^a}(\mathbf{z}_t^{des}, t, \mathbf{c}_{emb}) - \epsilon_{\theta^0}(\mathbf{z}_t^{des}, t, \mathbf{c}_{emb}), \tag{5}$$

$$\epsilon_{emb^*} = \epsilon_{emb} - \epsilon_{des}. \tag{6}$$

Note that ϵ_{emb} also contains design content information, and should not be pushed away from ϵ_{des}. Therefore, we use ϵ_{emb^*} to represent the subtracted embroidery features alone and push it away from the content features. We construct training batches, where each batch consists of the reference embroidery-design pair and a generated pair. For each batch, the final contrastive loss, denoted as $\mathcal{L}_{con}(\theta^e)$, is defined as follows:

$$\mathcal{L}_{con}(\theta^e) = -\log \frac{\exp(s(\epsilon^{ref}_{emb^*}, \epsilon^{gen}_{emb^*}))}{\exp(s(\epsilon^{ref}_{emb^*}, \epsilon^{gen}_{des})) + \exp(s(\epsilon^{ref}_{des}, \epsilon^{gen}_{emb^*}))}. \tag{7}$$

Here, we set the temperature τ to 1 and omit it for simplicity, and $s(\cdot,\cdot)$ denotes cosine similarity. The complementary embroidery images are from EmoLoRA generation and can have a good initial ϵ_{emb}^{gen}, while the design images are from the embroidery-to-design pipeline and make ϵ_{des}^{gen} unreasonable. To deal with this, we adopt a three-step iterative optimization. In Steps 1 and 2, we update θ^a and θ^e in Eqs. 2-3 as in the first stage, but on both the reference pair and the generated pair. In Step 3, we update θ^e with the contrastive loss $\mathcal{L}_{con}(\theta^e)$.

3.4 Model Inference

After the two-stage training of EmoLoRA, we apply model inference with image or text inputs, as in Alg. 2. For both settings, we only use the four embroidery blocks θ^e to update SDXL base model θ^0. For text inputs, which should include "*in [emb] style*", the model performs standard text-to-image synthesis. For image inputs, we adopt SDEdit [Meng et al. 2021] to first add noise to the input image, and then utilize the updated model to perform denoising under the guidance of text prompt "*in [emb] style*".

Similar to pair-wise data construction, we employ ControlNets to maintain the content from the input image, according to the style type. For styles such as flat embroidery, the boundaries between an image pair should be accurately aligned, we employ ControlNet-Tile and ControlNet-Canny, followed by a color correction module to enhance consistency with the input design. In this module, we first transfer the generated embroidery image to LAB space, then replace its A and B channels with the corresponding channels from the input design, and finally transfer the embroidery image back to RGB space. However, for embroideries with beads or sequins, we disable ControlNet-Canny and the color correction module to allow necessary modifications along the boundaries. Similar principles can be extended to other styles.

4 EXPERIMENT

We build a benchmark on embroidery customization to evaluate our method against prior art for fine-grained style customization, with ablation study to verify the efficacy of each component. In applications with customized embroideries, we explore the potential for transforming traditional embroidery workflows. Additionally, we extend our method to three additional style transfer tasks to illustrate its capability in decoupling style and content. For more implementation details, results, and discussions, please refer to the supplementary material.

4.1 Embroidery Dataset and Metrics

Dataset. We follow style transfer benchmarks [Chung et al. 2024; Deng et al. 2023, 2022] and build a dataset comprising 30 reference embroidery images and 50 test graphic design images. The reference set contains embroidery styles featuring various stitches and materials, including flat stich, towel stitch, beans, sequins, and more. For the test set, we use our embroidery-to-design module to generate the graphic design images with 50 additional embroidery images in any style, ensuring the test images are compatible with embroidery production. Additionally, we preset 20 text prompts for text-based generation. For each reference image, we evaluate the method across all test images and prompts.

ALGORITHM 2: EmoLoRA Inference

Input: Input text prompt p or design image $\hat{I}_{des}$, SDXL base model θ^0, trained embroidery blocks θ^e, ControlNet-Tile&Canny
Output: Embroidery image $\hat{I}_{emb}$

Update SDXL: $\theta \leftarrow \theta^0 + \theta^e$;
Update prompt: $p \leftarrow p$ + "in [emb] style";
if $\hat{I}_{des}$ *is not empty* **then**
  Employ ControlNet-Tile;
  if *Strict boundary alignment* **then**
    Employ ControlNet-Canny and color-correction;
  end
end
Generate image: $\hat{I}_{emb}$.

Metrics. For image-based customization, we adopt LPIPS [Zhang et al. 2018] and Histogram Loss [Afifi et al. 2021; Chung et al. 2024] to assess the preservation of design content and color. To evaluate embroidery style, we propose a metric named High-Frequency Ratio Difference (HFRD). Existing feature extractors such as VGG or CLIP mainly captures color, layout, or semantic features, while embroidery styles emphasize high-frequency structural textures, which makes existing metrics unsuitable for embroidery style evaluation. We propose HFRD to compute the absolute difference of the high-frequency energy ratio between the generated embroidery image and the reference. For text-based generation, we compute the CLIP-Score [Radford et al. 2021] between the generated image and the text prompt, to evaluate the level of semantic compliance. Additionally, we adopt Histogram Loss to assess the color difference between the generated embroidery and the reference, where a higher score means less similar in color and therefore better decoupling from reference content. The details and limitations of the metrics are discussed in the supplementary material, and we provide user studies in Sec. 4.4 to strengthen quantitative comparisons with different style transfer methods and our ablation variants.

4.2 Comparison on Embroidery Customization

Embroidery customization is a unique problem in contrast to general style transfer, as it redefines style and content. Elements like color must now be preserved as content rather than transferred as style, while high-frequency structural textures, often overlooked in artistic style transfer, become central to defining style. For embroidery customization, we focus on embroidery style similarity, design content preservation (for image inputs), decoupling of style from color and semantics (for text inputs). Qualitative results are in Fig. 5 and quantitative results are in Tab. 1.

We first compare with six prior methods in style transfer. On nine different embroidery styles in Fig. 5, our method exhibits high quality in fusion of the reference style and input image/text content. For DB-LoRA [Ryu 2022] and B-LoRA [Frenkel et al. 2025], we use their training approach but with our inference pipeline, to fully

Fig. 5. Comparison on one-shot embroidery customization using image/text inputs. The last row shows comparisons with two embroidery synthesis methods. **Ours (a)** and **Ours (b)** denote results using different reference embroideries. Please zoom in to examine detailed textures.

Table 1. Quantitative comparisons with style transfer methods. For image-based generation, we evaluate embroidery style quality (HFRD), design content preservation (LPIPS), and design color consistency (Histogram Loss). For text-based generation, we assess textual compliance (CLIP-Score) and reference color resemblance (Histogram Loss). The best results are highlighted in bold, and the second-best are underlined.

Metric	Ours	DB-LoRA	B-LoRA	InstantStyle	PairCustomization	StyleID	RB-Modulation
HFRD ↓ (embroidery style)	**6.50** ± 3.14	8.15 ± 4.22	<u>6.63</u> ± 2.25	12.41 ± 5.76	12.48 ± 5.17	21.66 ± 7.16	8.10 ± 3.97
LPIPS ↓ (design content)	<u>14.37</u> ± 9.66	14.54 ± 10.52	14.92 ± 8.03	**7.72** ± 7.47	22.14 ± 2.63	21.96 ± 2.59	65.18 ± 1.65
Histogram Loss ↓ (design color)	**26.59** ± 9.55	<u>28.62</u> ± 8.52	30.57 ± 7.97	32.23 ± 7.38	43.99 ± 1.99	45.75 ± 4.12	48.87 ± 1.61
CLIP-Score ↑ (text semantics)	<u>32.23</u> ± 0.61	30.94 ± 0.66	31.84 ± 0.33	25.14 ± 2.09	**32.47** ± 0.23	30.31 ± 0.92	30.04 ± 0.37
Histogram Loss ↑ (reference color)	**51.32** ± 6.82	43.89 ± 7.21	42.70 ± 6.09	33.64 ± 11.19	<u>50.49</u> ± 5.82	34.13 ± 9.89	35.48 ± 7.66

evaluate the style-content decoupling capability of our EmoLoRA. DB-LoRA can capture complete pearls or beads, while fails to fuse these structures with input content (Row 1, 3, 8) due to entanglement of style and reference content, and maintains the reference color or layout (Row 2, 6, 7, 9). B-LoRA uses a single *up_blocks.0.1* to capture style, possessing limited power in capturing embroidery structures and still presenting entanglement with the reference color. InstantStyle [Wang et al. 2024] injects reference features via pretrained IP-Adapter into the cross-attention of *up_blocks.0.1*, and thus captures even less embroidery features than B-LoRA. PairCustomization [Jones et al. 2024] adopts two LoRAs with orthogonal constraints to disentangle style and content, while freezing orthogonal matrices and training only one low-rank matrix also fails to capture complex embroidery textures. StyleID [Chung et al. 2024] blends style and content latent from DDIM inversion and leverages self-attention in later blocks, but produces blurry results due to entangled style and content features in pretrained self-attention. RB-Modulation [Rout et al. 2024] also fails to depict embroidery structures with an existing style descriptor, or preserve design content with the CLIP image encoder and attention feature aggregation. Moreover, we provide qualitative comparisons with Attention Distillation [Zhou et al. 2025] and Analogist [Gu et al. 2024] in the supplementary material, showing that approaches relying on pretrained attention features or textural descriptions fail to capture fine-grained styles such as embroidery.

In Tab. 1, Ours, DB-LoRA and B-LoRA have very similar results in HFRD and LPIPS, which is probably due to the limitations of current metrics in evaluating embroidery style and design content at a fine-grained level. InstantStyle achieves the best LPIPS as they mainly recover the input design. We attain the best score in Histogram Loss, demonstrating improved disentanglement from reference color. An ablation study on the color correction module and a discussion of metric limitations are included in the supplementary material. For text-based generation, our method achieves the highest CLIP-Score, indicating strong compliance with text prompts, and the highest Histogram Loss, reflecting minimal resemblance to the reference color.

Additionally, we compare to two methods for embroidery synthesis, both with three types of flat stitches. As in the last row of Fig. 5, we can generate highly realistic embroideries tailored to different references. MSEmbGAN [Hu et al. 2024] proposes a GAN-based approach and is bounded by the rendered training data. As their code and models are unavailable, we present a visual result generated with Wilcom EmbroideryStudio to approximate its upper bound, which exhibits a clear deficiency in photorealism. For the method of [Guan et al. 2021], we show their results in automatic mode (a) and long-stitch mode (b), which shows unnatural region division and rendering artifacts.

Table 2. User studies on overall embroidery quality, style consistency, and design preservation. The numbers represent the percentage of votes that these methods are preferred over our final model. For DB-LoRA, 19.44% in Quality means 19.44% of votes favor DB-LoRA's embroidery quality over ours, while 80.56% disagree.

Method	**Quality (%)**	**Style (%)**	**Design (%)**
DB-LoRA	19.44	24.05	27.63
B-LoRA	11.59	16.00	14.29
InstantStyle	10.68	2.50	31.58
PairCustomization	2.67	1.25	13.89
StyleID	1.30	0.00	5.13
RB-Modulation	13.16	3.75	0.00
2-Block Modulation	10.81	7.59	8.97
w/o Modulation	9.46	7.79	7.79
w/o Contrastive Learning	25.00	34.92	23.44

4.3 Ablation Study

We conduct ablation studies to analyze the efficacy of the components of our method. Specifically, we compare three variations: (1) **2-Block Modulation**, where we use the two LoRA blocks with the lowest average cosine similarity to capture embroidery style instead of four; (2) **w/o Modulation**, where we use all LoRA blocks instead of four to capture the embroidery style; and (3) **w/o Contrastive Learning**, where we adopt results from our first-stage training. The comparisons are shown in Fig. 4 (b). Using two blocks alone struggles to capture fine-grained embroidery structures, while using all blocks or omitting the second-stage contrastive learning fails to effectively decouple style from color and semantics, which causes unnatural fusion with the input design. We conduct user studies for quantitative evaluation due to the limitations of existing metrics, and provide additional ablation studies on using different blocks in the supplementary material.

4.4 User Study

To compensate for the misalignment of existing metrics and real objectives for embroidery customization, we conduct user studies to compare our methods to previous works and ablation variations.

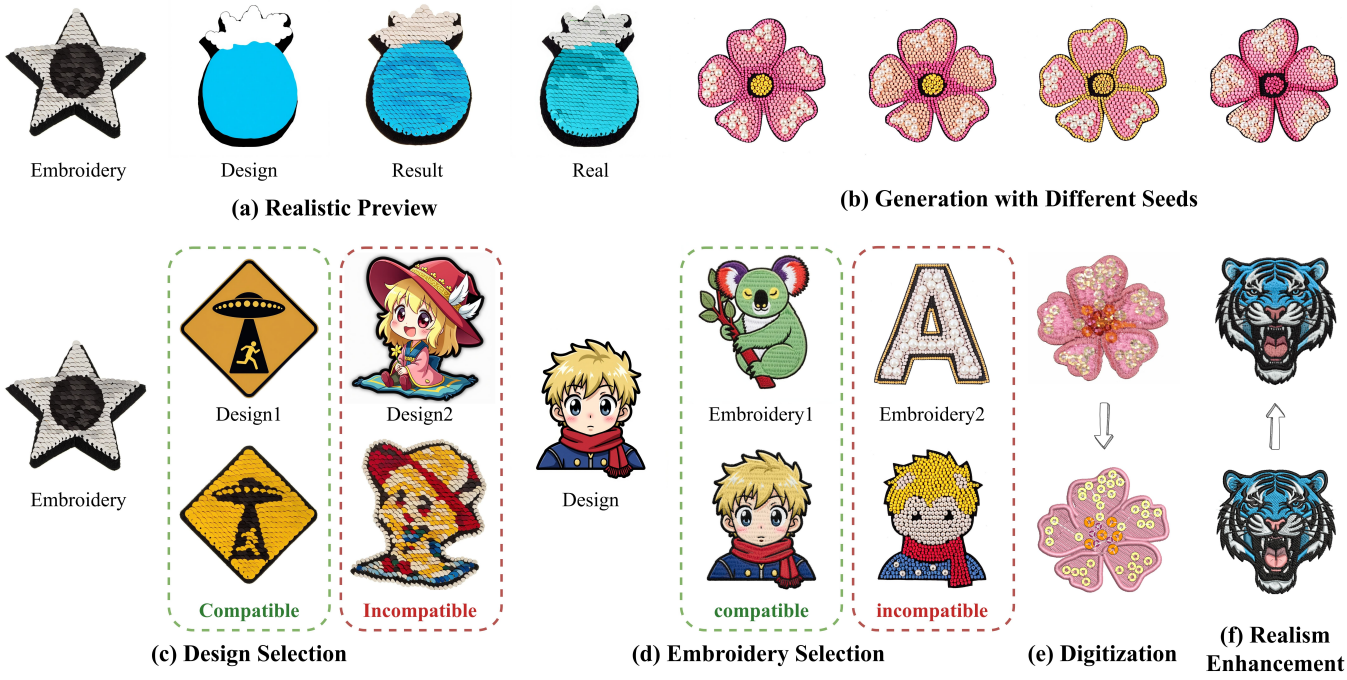


Fig. 6. Applications of embroidery customization. (a) The result from our generation is as realistic as real embroidery. (b) Our method can generate diverse results using different random seeds, offering more options for production. (c) Given a reference embroidery, our method can help identify compatible designs. (d) Given a design image, our method can suggest compatible embroidery styles. (e) Digitization by tracing a generated embroidery with Wilcom EmbroideryStudio. (f) Our method can enhance the realism of Wilcom EmbroideryStudio renderings.

Specifically, we follow similar setting as [Jones et al. 2024; Wang et al. 2023] and compare our method to others in pairs. For each user, we randomly sample 90 pairs, with each pair comprising a generated image from our method and one of another method. We provide the user study interface in the supplementary material. For each pair, users are asked to select their preferred option in terms of overall embroidery quality, style consistency with the reference, and design content preservation. Each question offers three choices: Method A, Method B, or Abstain.

We conducted an online questionnaire with 20 users, including two professionals in embroidery and 18 ordinary customers. Before answering, they viewed 10 reference embroidery images, followed by 10 generations with poor quality, inconsistent styles, or misaligned designs. All users completed the task within 30 minutes, though no time limit was set. The statistical results are presented in Tab. 2. Our method receives a clear preference over previous works, as well as over the ablation variations. Comparisons with other methods are based on valid votes. For Quality, Style, and Design, the valid ratios are 91.1%, 96.2%, and 94.2%, respectively, with each category comprising 1,800 votes in total.

4.5 Transformation to Embroidery Workflows

In this section, we illustrate the potential of our embroidery customization technique for transforming real-world embroidery workflows. Specifically, we first demonstrate its utility in enabling preview and presale, thereby bridging visual communication between producers and customers. We then showcase its role in fabrication support through embroidery digitization and visualization enhancement. Finally, we present more usage scenarios that highlight its capability to generate high-quality embroidery and design images.

Preview and Presale. With our generated embroidery previews, producer and consumer preferences can be better aligned, facilitating more effective presale decisions. As shown in Fig. 6 (a), we first verify that our generated results achieve realism comparable to real embroidery. Based on this, our method can then suggest compatible design patterns or suitable embroidery styles given a reference embroidery or a design image, as illustrated in Figs. 6 (c) and (d). In the supplementary material, we conduct user studies to evaluate whether participants can distinguish between real and generated embroidery images, and quantitatively assess how previews using our generated results influence their preferences.

Fabrication Support. Our embroidery customization technique supports the fabrication process in both digitization and visualization. Given a designated reference embroidery style and design image, our method can generate diverse outcomes (Fig. 6 (b)), thereby reducing iterative cycles of digitization and confirmation. Once a design is finalized, it can be digitized using Wilcom EmbroideryStudio to obtain a manufacturable file (Fig. 6 (e)). The tracing process involves color extraction, physical size alignment, layer-by-layer analysis, stitch filling, and decorative embellishment placement. Further details are provided in the supplementary document and

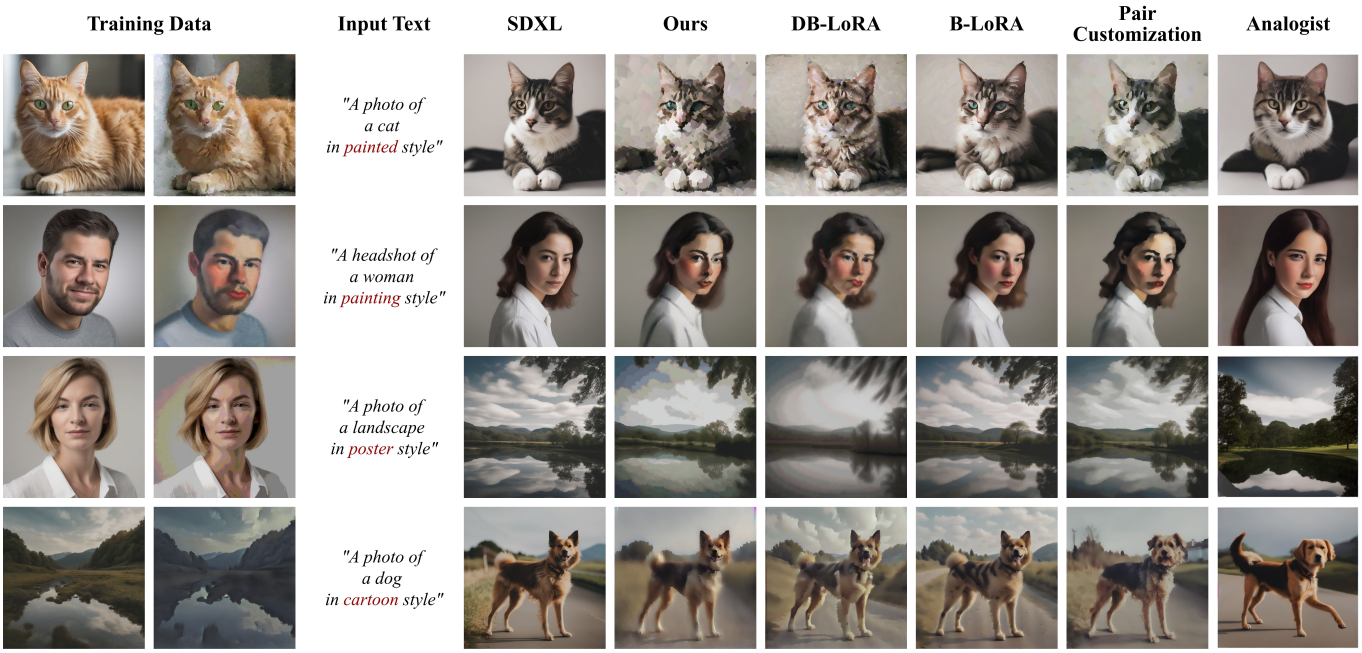


Fig. 7. Comparison on four examples of artistic style transfer. The training data (Row 2) © PairCustomization [Jones et al. 2024].

video. This example was completed within 10 minutes, although additional manual refinement is required for the final production file. In addition, our customization can serve as a realistic rendering module (Fig. 6 (f)), enhancing the realism of digitized embroidery and facilitating communication between producers and customers.

More Applications. Our generated embroidery can be overlaid onto garments, bags, or hats to provide more intuitive visual previews, through integration with ACE++ [Mao et al. 2025], as shown in Fig. 1. Moreover, the capability of our method in generating high-quality embroidery data helps address the challenge of data scarcity in this domain. Additionally, our embroidery-to-design module effectively recovers well-aligned design images from reference embroideries, enabling novel style synthesis with consistent designs. Additional visual results and implementation details are provided in the supplementary material.

4.6 Generalization to Other Styles

We evaluate our method across diverse styles to demonstrate its effectiveness in disentangling style and content. To this end, we compare against prior work on three tasks: artistic style transfer, sketch colorization, and appearance transfer. For each task, we construct a domain-specific data pair, followed by our standard block selection and contrastive LoRA learning. More results and implementation details are included in the supplementary material.

Photo to Artwork. We compare to PairCustomization, which also learns artistic style from a single image pair. Following their setup, we construct photo–artwork pairs using external stylization methods (Fig. 7). The objective is to generate stylized images while preserving SDXL-generated content from text inputs, thereby verifying style–content disentanglement during learning. We compare with LoRA-based methods using a timestep-controlled LoRA activation strategy during denoising, following PairCustomization. As in Fig. 7, DB-LoRA suffers from artifacts due to overfitting to training content, while B-LoRA yields weak stylization. Our method is comparable to PairCustomization on in-domain cases (Rows 1–2) and performs slightly better on cross-domain cases (Rows 3–4). Additional qualitative and quantitative analysis are provided in the supplementary material. These results confirm that our method achieves effective style–content disentanglement and offers a viable alternative for this task. We also evaluate Analogist with SDXL generation as its input for image analogy, and the results highlight its limitations in transferring such fine-grained styles using textual descriptions.

Sketch to Color. Our method can be applied to sketch colorization, through training on a color-Canny image pair to separate style (color and shading) from content (semantics and layout). In Fig. 8 (a), Ours achieves color consistency with the reference and effective content compliance. In contrast, DB-LoRA shows content entanglement (e.g., generating "standing in water" instead of "sitting on the floor" in the first row), while B-LoRA, InstantStyle, and PairCustomization exhibits noticeable color drift from the reference. ColorizeDiffusion [Yan et al. 2025] is trained on millions of colorization samples.

Appearance Transfer. We extend our method to appearance transfer by training on appearance–Canny pairs, where the style involves richer textures beyond color. At inference, we use HED [Xie and Tu 2015] maps of structure images and apply either style blocks as Ours (a), or all blocks as Ours (b), as shown in Fig. 8 (b). Ours (a) achieves better structural consistency through stronger appearance-structure

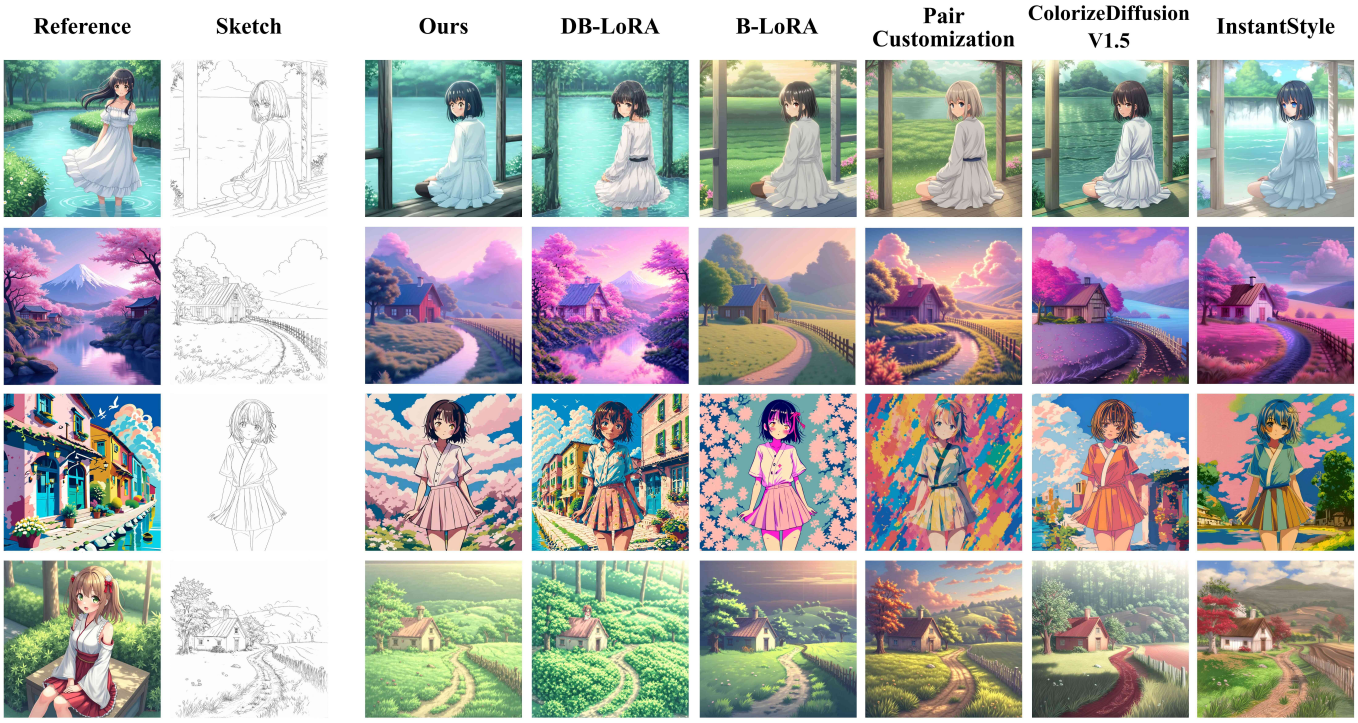


Fig. 8. Sketch colorization and appearance transfer. (a) Comparison on four examples of sketch-to-color style transfer. (b) Comparison on four examples of appearance transfer. For zebra, the content Canny map contains stripe patterns, leading to a decoupling between appearance and structure. Ours (a) uses only style blocks and captures distorted strips, while Ours (b) leverages all blocks and restores complete stripe structure.

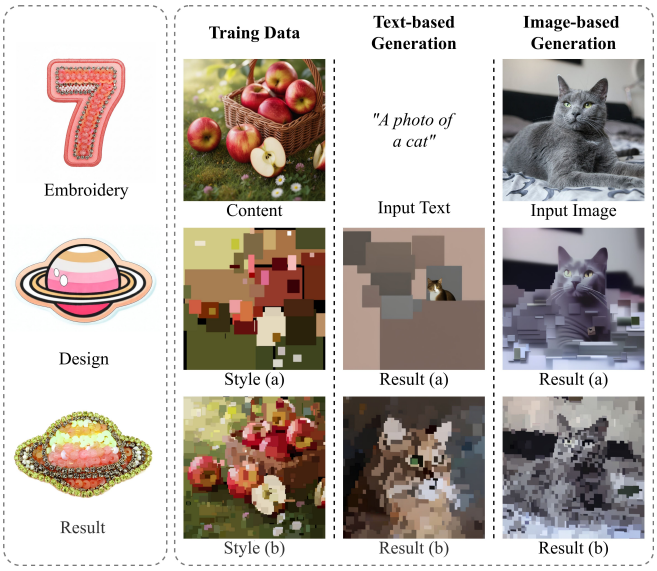


Fig. 9. Failure cases. The input image (Row 1, Column 4) © Milada Vigerova.

disentanglement, while Ours (b) preserves partial reference structure and fuses it compatibly with the input.

5 LIMITATION AND DISCUSSION

In this section, we present two representative failure cases, discuss the limitations of our method, and outline several future directions for supporting fabrication.

Failure Cases. Our method faces challenges with highly complex styles that combine multiple materials or overly abstract styles, as illustrated in Fig. 9. For **embroidery customization**, intricate reference styles—such as intersecting combinations of multiple bead types—often yield impractical customization results, while the reflective characteristics of rhinestones introduce additional imaging difficulties. Furthermore, localized style specification or editing is not yet supported. For **general style**, we employ the rectangle renderer from Stylized Neural Painting [Zou et al. 2021] to generate 8-bit artworks with two stroke configurations: (a) 50 strokes and (b) 550 strokes, as in Fig. 9 Style (a) and Style (b). Our method performs well when paired data are reasonably aligned, as in Style (b), but struggles with overly abstract styles such as Style (a), where block selection may fail to identify appropriate layers for modulation, leading to weak stylization and poor content blending.

Method Limitations. Our method comprises multiple stages: pair-wise data construction, network block selection, and two-stage training. We leverage SDXL for style–content disentanglement and SD3 as a plug-and-play module for design emulation, while the framework can be extended or unified with more powerful models exhibiting similar properties. We empirically select four blocks (e.g., 2, 3, 7, 8) to balance style completeness and style–content separation, though omitting blocks 1 and 11 can reduce color and appearance integrity in sketch colorization and appearance transfer. Our method could be enhanced with automatic block selection based on low cosine similarity and statistical constraints, and further refined to operate at a finer granularity through soft weighting rather than hard selection. In future work, we aim to reformulate fine-grained style customization within a meta-learning framework—e.g., treating the first training stage as meta-training to learn generalizable style disentanglement. Finally, when no reference embroidery is available, EmoLoRAs pretrained on tagged embroidery styles can serve as selectable modules, with text-only inputs mapped to style tags via LLMs, while combining style primitives from multiple references (e.g., chenille and sequin) remains an open direction.

Future Fabrication Support. To support future fabrication, we outline several directions for automatically generating production-ready files with rich structured information. A central step is defining a representation for primitive embroidery instructions (EmbIns)—including stitch coordinates, needle commands, and color sequences—analogous to SVG but with greater complexity. One promising avenue is the development of a differentiable rasterizer for EmbIns, akin to DiffVG [Li et al. 2020], enabling optimization-based generation of EmbIns from customized images. This approach could be extended to a unified framework for joint image and EmbIns generation, similar to VectorFusion [Jain et al. 2023]. Another direction is a multi-modal approach that tokenizes EmbIns for joint learning with text and images, as in OmniSVG [Yang et al. 2025], though this would require large-scale datasets for effective training.

6 CONCLUSION

In this paper, we address fine-grained style customization by introducing a contrastive learning framework that disentangles style and content from a single reference image, based on the classic concept of image analogy and leveraging decoupled representations from pretrained diffusion models. To capture fine-grained style features, we propose a two-stage contrastive LoRA modulation technique, EmoLoRA, which mitigates data scarcity through self-knowledge distillation. Our approach significantly outperforms existing methods in embroidery customization, with extensive analysis of its potential to transform real-world embroidery workflows. Moreover, it generalizes to three additional visual attribute transfer tasks, providing a new alternative to existing works.

For future work, we envision unifying pair-wise data generation with style customization into a single framework, enhancing block selection through automatic or soft-weighted strategies, and reformulating the two-stage learning as a meta-learning framework. Additionally, we outline directions for automatically generating production-ready embroidery files, drawing inspiration from SVG representations, including defining primitive embroidery instructions, developing differentiable rasterizers, and exploring joint learning of images, text, and embroidery instructions.

ACKNOWLEDGMENTS

This work was supported by Key R&D Program of Zhejiang (No. 2023C01047) and the Ningbo Major Special Projects of the "Science and Technology Innovation 2025" (Grant No. 2023Z143).

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown. 2021. Histogan: Controlling colors of gan-generated and real images via color histograms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7941–7950.

Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2024. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*. 1–12.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.

Xinling Chen, Michael McCool, Asanobu Kitamoto, and Stephen Mann. 2012. Embroidery modeling and rendering. In *Proceedings of Graphics Interface 2012*. 131–139.

Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. 2024. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8795–8805.

Dele Cui, Yun Sheng, and Guixu Zhang. 2017. Image-based embroidery modeling and rendering. *Computer Animation and Virtual Worlds* 28, 2 (2017), e1725.

Yusuf Dalva and Pinar Yanardag. 2024. Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24209–24218.

Yingying Deng, Xiangyu He, Fan Tang, and Weiming Dong. 2023. Z*: Zero-shot Style Transfer via Attention Rearrangement. *arXiv preprint arXiv:2311.16491* (2023).

Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. 2022. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11326–11336.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 8780–8794.

Alexei A Efros and William T Freeman. 2001. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. 341–346.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12873–12883.

Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. 2025. Implicit style-content separation using b-lora. In *European Conference on Computer Vision*. Springer, 181–198.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022).

Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2024. ReNoise: Real Image Inversion Through Iterative Noising. *arXiv preprint arXiv:2403.14602* (2024).

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2414–2423.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems* 27 (2014).

Zheng Gu, Shiyuan Yang, Jing Liao, Jing Huo, and Yang Gao. 2024. Analogist: Out-of-the-box visual in-context learning with image diffusion model. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–15.

Xinyang Guan, Likang Luo, Honglin Li, He Wang, Chen Liu, Su Wang, and Xiaogang Jin. 2021. Automatic embroidery texture synthesis for garment design and online display. *The Visual Computer* 37, 9 (2021), 2553–2565.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. 2024. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4775–4785.

Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. 2001. Image analogies. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. 327–340.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*. PMLR, 2790–2799.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

Xinrong Hu, Chen Yang, Fei Fang, Jin Huang, Ping Li, Bin ShengB, and Tong-Yee Lee. 2024. Msembgan: Multi-stitch embroidery synthesis via region-aware texture generation. *IEEE Transactions on Visualization and Computer Graphics* (2024).

Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1125–1134.

Ajay Jain, Amber Xie, and Pieter Abbeel. 2023. Vectorfusion: Text-to-svg by abstracting pixel-based diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1911–1920.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 694–711.

Maxwell Jones, Sheng-Yu Wang, Nupur Kumari, David Bau, and Jun-Yan Zhu. 2024. Customizing text-to-image models with a single image pair. In *SIGGRAPH Asia 2024 Conference Papers*. 1–13.

Tero Karras. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv preprint arXiv:1812.04948* (2019).

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8110–8119.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6007–6017.

Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1931–1941.

Gihyun Kwon and Jong Chul Ye. 2022. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264* (2022).

Tzu-Mao Li, Michal Lukáč, Michaël Gharbi, and Jonathan Ragan-Kelley. 2020. Differentiable vector graphics rasterization for editing and learning. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–15.

Wen Li, Muyuan Fang, Cheng Zou, Biao Gong, Ruobing Zheng, Meng Wang, Jingdong Chen, and Ming Yang. 2025. StyleTokenizer: Defining Image Style by a Single Instance for Controlling Diffusion Models. In *European Conference on Computer Vision*. Springer, 110–126.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal style transfer via feature transforms. *Advances in Neural Information Processing Systems* 30 (2017).

Zekun Li, Zhengyang Geng, Zhao Kang, Wenyu Chen, and Yibo Yang. 2022. Eliminating gradient conflict in reference-based line-art colorization. In *European Conference on Computer Vision*. Springer, 579–596.

Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. 2017. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088* (2017).

Jianghao Lin, Peng Du, Jiaqi Liu, Weite Li, Yong Yu, Weinan Zhang, and Yang Cao. 2025. Sell It Before You Make It: Revolutionizing E-Commerce with Personalized AI-Generated Items. *arXiv preprint arXiv:2503.22182* (2025).

Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. 2024. Towards Understanding Cross and Self-Attention in Stable Diffusion for Text-Guided Image Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7817–7826.

Chen Ma and Zhengxing Sun. 2022. Multilayered stitch generating for random-needle embroidery. *The Visual Computer* 38, 11 (2022), 3667–3679.

Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. 2025. ACE++: Instruction-Based Image Creation and Editing via Context-Aware Content Filling. *arXiv preprint arXiv:2501.02487* (2025).

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073* (2021).

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6038–6047.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4296–4304.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

Marion Nichols. 2012. *Encyclopedia of embroidery stitches, including crewel.* Courier Corporation.

Dae Young Park and Kwang Hee Lee. 2019. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5880–5888.

Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. 2020. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 319–345.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4195–4205.

Jessica Pile. 2018. *Fashion Embroidery: Embroidery Techniques and Inspiration for Haute-Couture Clothing*. Batsford Books.

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. *arXiv preprint arXiv:2307.01952* (2023).

Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. 2024. DEADiff: An Efficient Stylization Diffusion Model with Disentangled Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8693–8702.

Alec Radford, Jong Wook Kim, Chris Hallacy, A. 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*. 8748–8763.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

Litu Rout, Yujia Chen, Nataniel Ruiz, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. 2024. RB-Modulation: Training-Free Personalization of Diffusion Models using Stochastic Optimal Control. *arXiv preprint arXiv:2405.17401* (2024).

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

Simo Ryu. 2022. *Low-rank adaptation for fast text-to-image diffusion fine-tuning*. https://github.com/cloneofsimo/lora

Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. 2025. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*. Springer, 422–438.

Qiqi Shen, Dele Cui, Yun Sheng, and Guixu Zhang. 2017. Illumination-preserving embroidery simulation for non-photorealistic rendering. In *MultiMedia Modeling: 23rd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part II 23*. Springer, 233–244.

SmilingWolf. 2023. wd-convnext-tagger-v3. https://huggingface.co/SmilingWolf/wd-convnext-tagger-v3.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, 2256–2265.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. 2024. Measuring Style Similarity in Diffusion Models. *arXiv preprint arXiv:2404.01292* (2024).

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

Adéla Šubrtová, Michal Lukáč, Jan Čech, David Futschik, Eli Shechtman, and Daniel Sỳkora. 2023. Diffusion image analogies. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–10.

Luming Tang, Nataniel Ruiz, Qinghao Chu, Yuanzhen Li, Aleksander Holynski, David E Jacobs, Bharath Hariharan, Yael Pritch, Neal Wadhwa, Kfir Aberman, et al. 2024. Realfill: Reference-driven generation for authentic image completion. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–12.

Joshua Tenenbaum and William Freeman. 1996. Separating style and content. *Advances in Neural Information Processing Systems* 9 (1996).

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. 2023. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2022. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10748–10757.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1921–1930.

Dani Valevski, Matan Kalman, Eyal Molad, Eyal Segalis, Yossi Matias, and Yaniv Leviathan. 2023. Unitune: Text-driven image editing by fine tuning a diffusion model on a single image. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–10.

Haofan Wang, Matteo Spinelli, 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).

Zhizhong Wang, Lei Zhao, and Wei Xing. 2023. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7677–7689.

Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. 2021. Styleformer: Real-time arbitrary style transfer via parametric style composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14618–14627.

Saining Xie and Zhuowen Tu. 2015. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*. 1395–1403.

Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. 2024. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766* (2024).

Dingkun Yan, Xinrui Wang, Zhuoru Li, Suguru Saito, Yusuke Iwasawa, Yutaka Matsuo, and Jiaxian Guo. 2025. Image Referenced Sketch Colorization Based on Animation Creation Workflow. *arXiv preprint arXiv:2502.19937* (2025).

Chen Yang, Xinrong Hu, Yangjun Ou, Saishang Zhong, Tao Peng, Lei Zhu, Ping Li, and Bin Sheng. 2022. Unsupervised Embroidery Generation Using Embroidery Channel Attention. In *Proceedings of the 18th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*. 1–8.

Kewei Yang and Zhengxing Sun. 2018. Paint with stitches: a style definition and image-based rendering method for random-needle embroidery. *Multimedia Tools and Applications* 77 (2018), 12259–12292.

Kewei Yang, Jie Zhou, Zhengxing Sun, and Yi Li. 2012. Image-based irregular needling embroidery rendering. In *proceedings of the 5th International Symposium on Visual Information Communication and Interaction*. 87–94.

Serin Yang, Hyunmin Hwang, and Jong Chul Ye. 2023. Zero-shot contrastive loss for text-guided diffusion image style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22873–22882.

Yiying Yang, Wei Cheng, Sijin Chen, Xianfang Zeng, Fukun Yin, Jiaxu Zhang, Liao Wang, Gang Yu, Xingjun Ma, and Yu-Gang Jiang. 2025. Omnisvg: A unified scalable vector graphics generation model. *arXiv preprint arXiv:2504.06263* (2025).

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. *arXiv preprint arXiv:2308.06721* (2023).

Jingwen Ye, Yixin Ji, Jie Song, Zunlei Feng, and Mingli Song. 2021. Towards End-to-End Embroidery Style Generation: A Paired Dataset and Benchmark. In *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part IV 4*. Springer, 201–213.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023c. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

Qian Zhang, Bo Wang, Wei Wen, Hai Li, and Junhui Liu. 2021. Line art correlation matching feature transfer network for automatic animation colorization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3872–3881.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*. 586–595.

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 Transactions on Graphics (TOG)* 42, 6 (2023), 1–14.

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023b. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10146–10156.

Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. 2022. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–8.

Liu Zhenyuan, Michal Piovarči, Christian Hafner, Raphaël Charrondière, and Bernd Bickel. 2023. Directionality-Aware Design of Embroidery Patterns. *Computer Graphics Forum* 42, 2 (2023). https://doi.org/10.1111/cgf.14770

Yang Zhou, Xu Gao, Zichong Chen, and Hui Huang. 2025. Attention distillation: A unified approach to visual characteristics transfer. *arXiv preprint arXiv:2502.20235* (2025).

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2223–2232.

Zhengxia Zou, Tianyang Shi, Shuang Qiu, Yi Yuan, and Zhenwei Shi. 2021. Stylized neural painting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15689–15698.