CoStitch: Disentangled Color Style and Stitch Texture Generation for Controllable Embroidery Synthesis

Anonymous authors
Paper under double-blind review

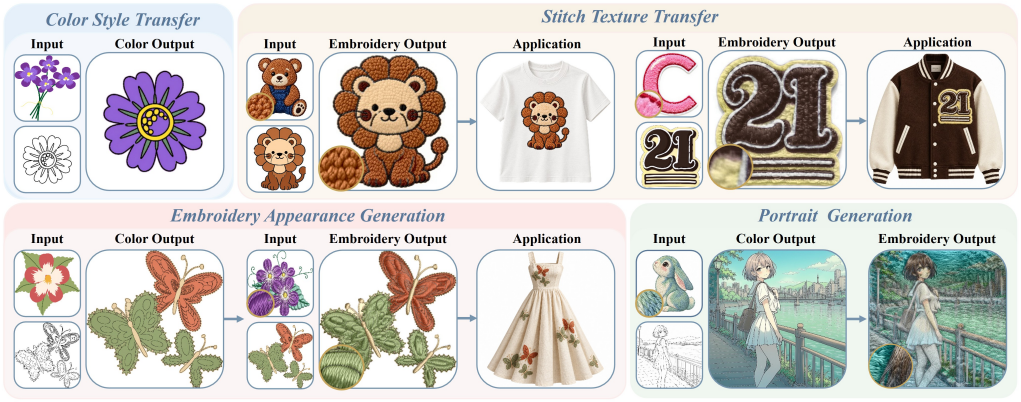


Figure 1: Overview of CoStitch for color and stitch texture transfer. Their combination enables coherent embroidery appearance generation across diverse applications and complex scenes.

Abstract

Embroidery image synthesis aims to generate realistic embroidery images with specified color styles and stitch textures while preserving target pattern structure. However, existing methods typically encode color style and stitch texture within a unified appearance representation, hindering their disentanglement and independent reuse on novel patterns. To address this issue, we propose CoStitch, a framework for disentangled color style and stitch texture generation. CoStitch exploits three progressively enriched image states to provide differentiated supervision for the two attributes. Specifically, CoStitch constructs an aligned triplet comprising a sketch, a color design image, and an embroidery image, where the color design image serves as an intermediate state that distinguishes color-related variation from stitch-texture variation. CoStitch further performs joint image-text response analysis to identify color-sensitive and texture-sensitive blocks, defining attribute-specific learning subsets for Color LoRA and Stitch LoRA to reduce cross-attribute interference. The three image states are then reconstructed successively with corresponding LoRA updates, progressively learning color style and fine-grained stitch texture. Finally, texture enhancement stops sketch reconstruction and further refines stitch direction, length, density, and arrangement while preserving the learned color style. Experiments demonstrate that CoStitch preserves target pattern structure while enabling color style and stitch texture to be reused independently and flexibly combined across different references.

1 Introduction

Embroidery is a traditional textile art in which colored threads are guided by a needle along design contours to form patterns through variations in thread color and stitch direction, length, and arrangement. In high-end fashion, embroidery is widely used in garments, bags, and accessories as

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

both decoration and a design language for brand identity. Designers combine pattern structure, color style, and stitch texture to create embroidery with distinctive craftsmanship and brand characteristics. However, their diverse combinations make conventional design highly dependent on repeated manual adjustment. Deep generative models have recently been applied to automated fashion and image generation (Yan et al., 2024; Ho et al., 2020; Rombach et al., 2022). Nevertheless, preserving the target pattern structure while reproducing the reference color style and stitch texture remains challenging for embroidery synthesis. Figure 1 illustrates the main capabilities of CoStitch.

A central objective of embroidery synthesis is to reuse color style and stitch texture while preserving target pattern structure. Existing studies employ image style transfer to extract and transfer reference characteristics. Classical methods represent reference style using statistics such as Gram matrices (Gatys et al., 2016), feature statistics (Huang & Belongie, 2017), or covariances (Li et al., 2017). Although effective for global color and texture transfer, they remain limited in modeling stitch direction, continuity, and density. Subsequent methods improve local texture transfer through multi-scale interaction (Sheng et al., 2018), non-local matching (Deng et al., 2020), attention mechanisms (Park & Lee, 2019), and transformer architectures (Deng et al., 2022). However, they still treat reference appearance as a unified style, limiting separate specification of color style and stitch texture. Recent methods improve style fidelity and generalization (Huang et al., 2023; Tang et al., 2023; Zheng et al., 2024), but still model overall style without distinguishing these attributes. Consequently, the two attributes remain difficult to extract and reuse independently from different references.

Pretrained diffusion models enable visual characteristics to be learned from a limited number of reference images. Existing methods reproduce reference appearance through model fine-tuning (Ruiz et al., 2023), partial parameter updating (Kumari et al., 2023), or low-rank adaptation (Hu et al., 2022). However, these approaches typically encode pattern structure, color style, and stitch texture within a shared appearance representation. Another line of work reduces content-style coupling through self-attention feature injection (Chung et al., 2024), network block selection (Frenkel et al., 2024), or disentangled parameter learning (Jones et al., 2024; Liu et al., 2025a). Nevertheless, these methods mainly disentangle content from overall style. For embroidery synthesis, separating the learning of color style and stitch texture requires attribute-specific supervision and parameter optimization to reduce interference between the two attributes.

To address these issues, we propose CoStitch, a framework for disentangled color style and stitch texture generation. CoStitch employs an aligned triplet of sketch, color design, and embroidery images, where the color design image serves as an intermediate state to provide differentiated supervision for color-related and stitch-texture variations. Joint image-text response analysis identifies color-sensitive and texture-sensitive blocks, defining attribute-specific parameter subsets for Color LoRA and Stitch LoRA. The three image states are successively reconstructed with corresponding LoRA updates to progressively learn color style and fine-grained stitch texture. Texture enhancement further refines stitch characteristics while preserving the learned color representation. CoStitch thereby supports the independent reuse and cross-reference combination of color style and stitch texture while preserving target structure. Our main contributions are as follows:

- We propose a joint image-text attribute-sensitive block localization method that combines image-state and textual-condition responses across denoising intervals to identify color-sensitive and texture-sensitive blocks.
- We propose a collaborative learning method that exploits differentiated supervision from three image states and performs successive reconstruction with corresponding LoRA updates to separate color style and stitch texture learning.
- We introduce a texture enhancement strategy that stops sketch reconstruction and disables Color LoRA, while successively reconstructing color design and embroidery images to refine fine-grained stitch texture while preserving color style.

2 Related Work

2.1 Embroidery Image Generation

Early embroidery image generation relied on CAD, stitch modeling, and texture rendering. Chen et al. (2012) combined stitch arrangements with a yarn illumination model to render embroidery appearance; Yang et al. (2012; 2016) simulated random-needle embroidery using direction fields and

layered rendering; Takahashi & Fukusato (2018) developed an interactive CAD system; and Guan et al. (2021) synthesized embroidery textures through region partitioning and texture mapping. More recently, MSEmbGAN (Hu et al., 2025) generated multi-stitch embroidery images, while Ma et al. (2025) learned fine-grained embroidery appearance from a reference embroidery image. However, these methods generally model color style and stitch texture jointly as part of the overall embroidery appearance, making the two attributes difficult to reuse from different references.

2.2 Traditional Neural Image Style Transfer

Image style transfer aims to transfer reference appearance to target images while preserving semantic and structural information. Early methods represent style using feature correlations and statistics (Gatys et al., 2016; Huang & Belongie, 2017; Li et al., 2017), while later approaches improve texture modeling through attention, feature matching, and style composition (Liu et al., 2021; Wu et al., 2021; Zhang et al., 2022). Subsequent studies enhance transfer quality and structure preservation (Xie et al., 2022; Wen et al., 2023; Hong et al., 2023). However, these methods encode color and texture within a unified style representation. Recent approaches improve visual fidelity and efficiency (Zhu et al., 2023; Zhang et al., 2023; Liu et al., 2025b), but remain focused on content-style separation rather than the finer-grained distinction between color style and stitch texture.

2.3 Diffusion-Based Style Transfer and Disentanglement Models

Pretrained diffusion models enable visual characteristics to be learned from limited references through full-model fine-tuning, embedding optimization, or parameter-efficient adaptation (Ruiz et al., 2023; Gal et al., 2023; Kumari et al., 2023; Hu et al., 2022). However, these methods often provide limited control over individual appearance attributes. StyleInjection (Chung et al., 2024), StyleAligned (Hertz et al., 2024), InstantStyle (Wang et al., 2024), ArtAdapter (Chen et al., 2024), and Attention Distillation (Zhou et al., 2025) improve reference appearance transfer but still encode reference appearance within a unified representation. B-LoRA (Frenkel et al., 2024), PairCustomization (Jones et al., 2024), ZipLoRA (Shah et al., 2024), and UnZipLoRA (Liu et al., 2025a) further explore content-style separation. Despite reducing content-style coupling, these methods mainly disentangle content from overall style rather than color style from stitch texture, while finer-grained attribute supervision and attribute-specific parameter optimization remain insufficiently explored.

3 Method

This section presents CoStitch, which leverages three progressively enriched image states to provide differentiated supervision for color style and fine-grained stitch texture learning while preserving target structure. As illustrated in Figure 2, CoStitch comprises attribute-sensitive block localization, successive reconstruction, and stitch texture enhancement.

3.1 Problem Definition

In embroidery design, pattern structure, color style, and stitch texture jointly determine visual appearance. However, color style and stitch texture are coupled with pattern structure: thread colors occupy specific regions, while stitch directions and arrangements depend on local contours, making the two attributes difficult to learn and transfer separately. Therefore, CoStitch aims to disentangle the learning of color style and stitch texture while retaining their dependence on pattern structure. Here, disentanglement refers to separating the supervision and parameter optimization of the two attributes to support their independent reuse, rather than enforcing statistical independence between their latent representations.

Given an embroidery image I_e, HED (Xie & Tu, 2015) extracts a sketch I_s that characterizes pattern structure, while Wilcom (Sofronova & Angelova, 2019) constructs a stitch-free color design image I_d that preserves color information without stitch texture. As an intermediate state, I_d distinguishes color-related variation between I_s and I_d from stitch-texture variation between I_d and I_e, thereby providing differentiated supervision for the two attributes. The triplet (I_s, I_d, I_e) is associated with text descriptions (T_s, T_d, T_e). CoStitch successively reconstructs the three image states under their

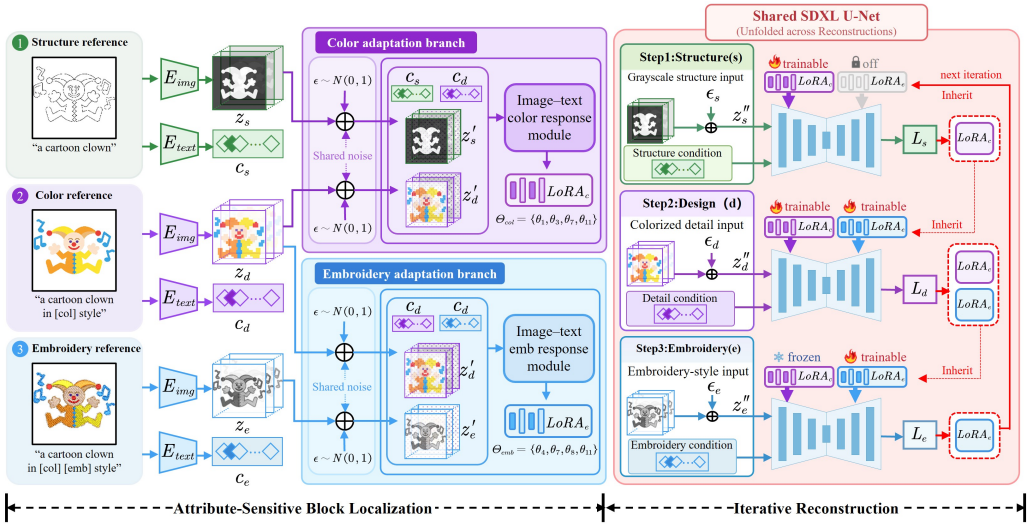


Figure 2: Overall framework of CoStitch. Attribute-sensitive block localization identifies the learning subsets for Color LoRA and Stitch LoRA, followed by successive reconstruction of the three image states with corresponding LoRA updates.

corresponding textual conditions as $\mathcal{G}((I_s, T_s), (I_d, T_d), (I_e, T_e)) \rightarrow (\hat{I}_s, \hat{I}_d, \hat{I}_e)$, where $\hat{I}_s$, $\hat{I}_d$, and $\hat{I}_e$ denote the reconstructed sketch, color design image, and embroidery image, respectively.

3.2 Color Style Learning and Color-Sensitive Block Selection

Color Encoding. To learn the dominant color style represented by the color design image, we encode the sketch I_s and color design image I_d into the latent space. Specifically, a pretrained VAE encoder (Rombach et al., 2022) maps I_s and I_d to latent representations, yielding $z_s = E_{\text{img}}(I_s)$ and $z_d = E_{\text{img}}(I_d)$. The dual text encoder of SDXL (Podell et al., 2024) encodes their text descriptions into conditioning features $c_s = E_{\text{text}}(T_s)$ and $c_d = E_{\text{text}}(T_d)$.

Color-Sensitive Block Selection. Let the N attention blocks in the U-Net constitute the candidate set, with their corresponding LoRA parameters denoted by $\Theta = \{\theta_i\}_{i=1}^{N}$, where $N = 11$. Since candidate blocks exhibit different responses to color-related variations, we identify the blocks with the strongest color sensitivity and use the resulting subset for attribute-specific Color LoRA optimization during color design reconstruction. This attribute-specific parameter selection reduces interference with subsequent stitch texture learning.

The variation between I_s and I_d provides the primary image-level evidence for color style. We therefore jointly analyze image-state and textual-condition responses to characterize the color sensitivity of each candidate block. As illustrated in Figure 3, the image-state response is evaluated by varying the latent state from z_s to z_d while fixing c_s, whereas the textual-condition response is evaluated by varying c_s to c_d while fixing z_s. The same Gaussian noise is applied to z_s and z_d at the same timestep, yielding z'_s and z'_d. For each candidate block, the two responses are defined as

$$r_1 = \mathcal{D}\big(H(z'_s, c_s), H(z'_d, c_s)\big), \qquad r_2 = \mathcal{D}\big(H(z'_s, c_s), H(z'_s, c_d)\big), \tag{1}$$

where $H(\cdot)$ denotes the output feature of the candidate block and $\mathcal{D}(\cdot)$ denotes the cosine distance between flattened features. Here, r_1 measures the response to image-state variation, while r_2 measures the response to textual-condition variation.

Because block responses vary throughout denoising, we compute r_1 and r_2 along a 50-step trajectory divided into 10 consecutive intervals. Within each interval, the responses are averaged and normalized to obtain $\bar{r}_1$ and $\bar{r}_2$. Their geometric mean $u = \sqrt{\bar{r}_1 \bar{r}_2}$ produces a high fused response only when both responses are strong. Since denoising intervals contribute differently to color sensitivity,

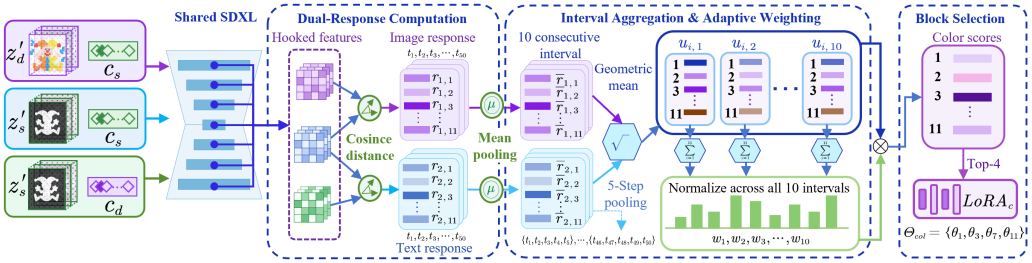


Figure 3: Overview of the joint image-text response analysis for color-sensitive block localization across denoising intervals.

we assign adaptive interval weights and compute the final score as

$$w_k = \frac{\sum_{i=1}^{N} u_{i,k}}{\sum_{j=1}^{10}\sum_{i=1}^{N} u_{i,j} + \delta}, \qquad s_i = \sum_{k=1}^{10} w_k u_{i,k}, \tag{2}$$

where $u_{i,k}$ is the fused response of block i in interval k, w_k is the interval weight, and s_i is the final color-sensitivity score. The four highest-scoring blocks define the Color LoRA learning subset $\Theta_{\text{col}} = \{\theta_1, \theta_3, \theta_7, \theta_{11}\}$.

Color Style Learning. After obtaining Θ_{col}, we incorporate Color LoRA (LoRA_c) and Stitch LoRA (LoRA_e) into the candidate attention blocks. The sketch and color design image provide supervision for pattern structure and color style, respectively. For z_s and z_d, a diffusion timestep and Gaussian noise are independently sampled to obtain z_s'' and z_d'', with the corresponding noise-prediction losses

$$\mathcal{L}_s = \|\epsilon_s - \epsilon_\phi(z_s'', t, c_s)\|_2^2, \qquad \mathcal{L}_d = \|\epsilon_d - \epsilon_\phi(z_d'', t, c_d)\|_2^2. \tag{3}$$

Here, ϵ_s and ϵ_d denote the sampled Gaussian noise, and ϵ_ϕ denotes the frozen U-Net predictor. $\mathcal{L}_s$ first updates all Color LoRA parameters under structural supervision. Subsequently, $\mathcal{L}_d$ updates Color LoRA within Θ_{col} together with all Stitch LoRA parameters. Color LoRA learns the color style, while Stitch LoRA establishes an intermediate representation from the color design state that is subsequently specialized toward stitch texture during embroidery reconstruction.

3.3 Stitch Texture Learning and Texture-Sensitive Block Selection

Texture Encoding and Sensitive Block Selection. Building upon the intermediate representation established from the color design state, we specialize Stitch LoRA toward fine-grained stitch texture using the image-state variation between I_d and I_e. The embroidery image I_e and its textual description T_e are encoded as $z_e = E_{\text{img}}(I_e)$ and $c_e = E_{\text{text}}(T_e)$, respectively. We then apply the joint image-text block localization method described above to the color design and embroidery images. To ensure comparable block responses across image states, the same Gaussian noise is applied to z_d and z_e at the same timestep, yielding z_d' and z_e'. For each candidate block, the image-state and textual-condition responses are defined as

$$q_1 = \mathcal{D}\big(H(z_d', c_d), H(z_e', c_d)\big), \qquad q_2 = \mathcal{D}\big(H(z_d', c_d), H(z_d', c_e)\big). \tag{4}$$

Here, q_1 measures the block response to image-state variation between I_d and I_e, whereas q_2 measures the block response to textual-condition variation from c_d to c_e. The response fusion and denoising-interval weighting described in Section 3.2 are then applied to compute the texture-sensitivity score for each candidate block. The four highest-scoring blocks are selected as the texture-sensitive blocks, defining the Stitch LoRA learning subset $\Theta_{\text{emb}} = \{\theta_4, \theta_7, \theta_8, \theta_{11}\}$.

Stitch Texture Learning. After obtaining Θ_{emb}, we reconstruct the embroidery image to learn fine-grained stitch texture. A diffusion timestep t and Gaussian noise ϵ_e are sampled for z_e, yielding z_e'', with the corresponding noise-prediction loss

$$\mathcal{L}_e = \|\epsilon_e - \epsilon_\phi(z_e'', t, c_e)\|_2^2. \tag{5}$$

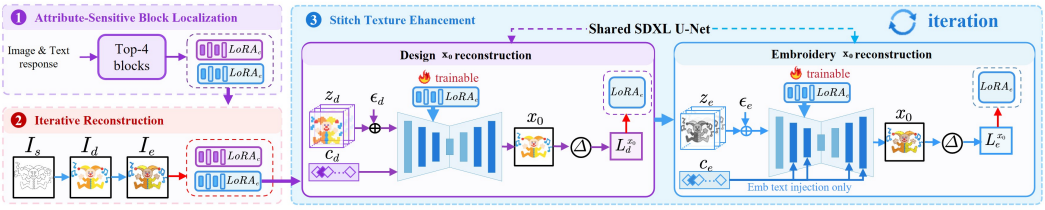
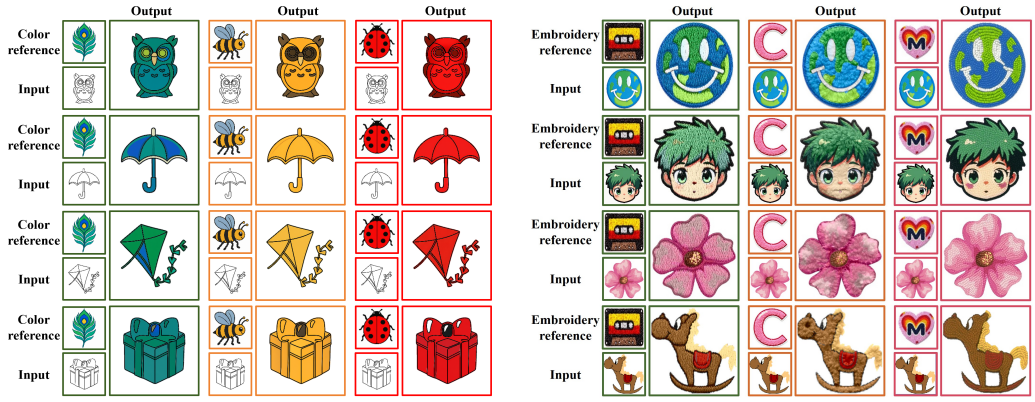


Figure 4: Stitch texture enhancement in CoStitch. The left summarizes attribute-sensitive block localization and successive reconstruction, while the right shows texture enhancement, where Color LoRA is disabled and Stitch LoRA is further optimized through successive reconstruction of the color design and embroidery images.



(a) Color style generation.

(b) Stitch texture transfer.

Figure 5: Individual transfer results of CoStitch. (a) Color style transfer while preserving target structure. (b) Fine-grained stitch texture transfer while preserving target structure and color style.

During embroidery reconstruction, the pretrained U-Net remains frozen. Color LoRA participates in the forward pass but is not updated, preserving the color style learned from the preceding color design reconstruction. The loss $\mathcal{L}_e$ updates only Stitch LoRA within Θ_{emb}, restricting stitch texture learning to the selected parameter subset and reducing cross-attribute interference.

3.4 Texture Enhancement Strategy

To further enhance fine-grained stitch texture while preserving the learned color style, we stop sketch reconstruction and disable Color LoRA during texture enhancement. As illustrated in Figure 4, the color design image and embroidery image are successively reconstructed based on the predicted x_0, with one Stitch LoRA update performed after each reconstruction. Compared with noise-prediction supervision, x_0-based reconstruction directly constrains the reconstructed latent states, thereby providing more explicit supervision for local stitch structures. For the color design image, all candidate blocks use the textual condition c_d. During embroidery reconstruction, however, the textual conditions are assigned according to block type: texture-sensitive blocks use c_e, while the remaining candidate blocks retain c_d. This block-specific conditioning restricts stitch texture guidance to the texture-sensitive blocks while preserving the color representation established by the color design image. Under this conditioning scheme, the predicted x_0 states for the color design and embroidery images are recovered as

$$\hat{z}_d = \frac{z''_d - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\phi(z''_d, t, c_d)}{\sqrt{\bar{\alpha}_t}}, \qquad \hat{z}_e = \frac{z''_e - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\phi(z''_e, t, c_d, c_e)}{\sqrt{\bar{\alpha}_t}}, \tag{6}$$

where $\epsilon_\phi(z''_e, t, c_d, c_e)$ denotes the noise prediction under the block-specific conditioning described above, and $\bar{\alpha}_t$ denotes the cumulative noise-schedule coefficient. Using z_d and z_e as reconstruction

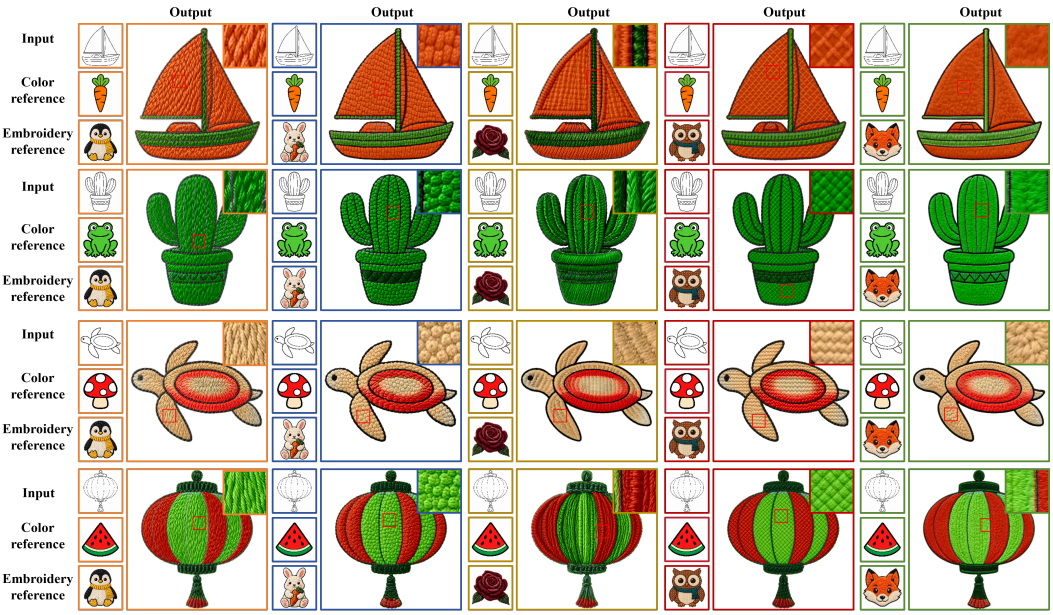


Figure 6: Cross-reference collaborative generation using independently specified color style and stitch texture. Each group shows a target sketch, color reference, texture reference, intermediate color design image, and final embroidery result.

targets, the corresponding losses are

$$\mathcal{L}_d^{x_0} = \|\hat{z}_d - z_d\|_2^2, \qquad \mathcal{L}_e^{x_0} = \|\hat{z}_e - z_e\|_2^2. \tag{7}$$

The loss $\mathcal{L}_d^{x_0}$ first updates all Stitch LoRA parameters. Subsequently, the updated parameters are inherited by the embroidery reconstruction, where $\mathcal{L}_e^{x_0}$ updates only the Stitch LoRA parameters within Θ_{emb}. Through this successive optimization, Stitch LoRA further refines stitch direction, length, density, and local arrangement while preserving the learned color style. Throughout texture enhancement, the pretrained U-Net remains frozen and Color LoRA remains disabled.

4 Experiments

4.1 Independent and Collaborative Generation

CoStitch is customized using a single reference triplet and evaluated on unseen target patterns. Figure 5 shows that the learned color style and stitch texture can be independently reused while preserving target structure. Figure 6 further combines the two attributes from different references, demonstrating flexible cross-reference generation.

4.2 Comparison with Existing Methods

4.2.1 Evaluation Metrics and Baselines

We evaluate structure preservation using Silhouette Dice (Dice, 1945), color similarity using Palette EMD (Rubner et al., 2000) and Palette ΔE_{00} (Sharma et al., 2005), and stitch texture similarity using GLCM-TD (Haralick et al., 1973), HFRD (Ma et al., 2025), and SIFID (Shaham et al., 2019). Higher values indicate better performance for Silhouette Dice, whereas lower values indicate better performance for all other metrics.

We compare CoStitch with EmoLoRA (Ma et al., 2025), B-LoRA (Frenkel et al., 2024), DB-LoRA (Ruiz et al., 2023; Hu et al., 2022), PairCustomization (Jones et al., 2024), InstantStyle (Wang et al., 2024), StyleID (Chung et al., 2024), and RB-Modulation (Rout et al., 2025), following their original input settings. All metrics are computed within the foreground regions, using the target

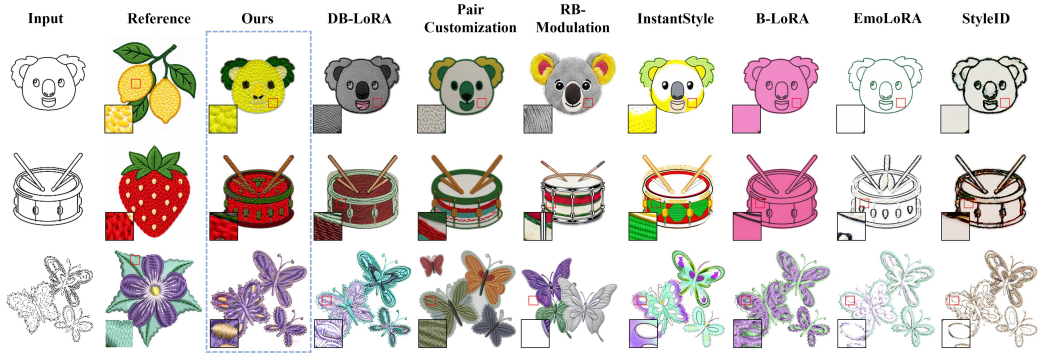


Figure 7: Qualitative comparison with existing methods.

Table 1: Quantitative comparison with existing methods.

Dimension	Metric	Ours	EmoLoRA	B-LoRA	DB-LoRA	Pair	InstantStyle	StyleID	RBM
Structure	Silhouette Dice ↑	0.9593	0.9518	0.7114	0.5743	0.5788	0.6668	**0.9784**	0.7392
Texture	HFRD ↓	**0.0048**	0.0055	0.0094	0.0333	0.0175	0.0138	0.0192	0.0107
	GLCM-TD ↓	**0.0508**	0.1534	0.2798	0.1052	0.1244	0.1919	0.2901	0.0763
	SIFID ↓	**14.1072**	49.9948	113.1927	124.1957	130.5730	62.4168	47.0082	17.7102
Color	Palette EMD ↓	**30.0398**	42.5897	49.7610	37.3731	36.9678	36.1494	41.9923	33.2912
	Palette ΔE_{00} ↓	10.5836	16.9395	19.6832	12.8681	13.0586	**10.5558**	12.6144	11.8794

sketch as the structural reference and the reference embroidery image for evaluating color style and stitch texture.

4.2.2 Qualitative and Quantitative Comparison

Figure 7 presents qualitative comparisons across different references and target patterns. DB-LoRA and PairCustomization exhibit color deviations or blurred stitch textures, while RB-Modulation and InstantStyle preserve target structure but show weaker stitch continuity and local arrangement. B-LoRA and StyleID exhibit color or texture mismatches, whereas the reproduced first-stage EmoLoRA result mainly preserves target contours. In comparison, CoStitch preserves the overall target structure while reproducing the dominant color appearance and fine-grained stitch patterns of the reference. As shown in Table 1, CoStitch achieves a Silhouette Dice of 0.9593, slightly lower than StyleID at 0.9784. This difference is associated with fine-grained stitches near target boundaries, which introduce slight variations in the foreground silhouette. For color appearance, CoStitch achieves the best Palette EMD of 30.0398, while its Palette ΔE_{00} of 10.5836 is only 0.0278 higher than the best result. Since CoStitch reproduces dominant colors and color relationships rather than establishing region-wise color correspondence, local color differences may affect the ΔE_{00} score. For stitch texture, CoStitch achieves the best HFRD, GLCM-TD, and SIFID scores of 0.0048, 0.0508, and 14.1072, improving over the second-best results by 12.7%, 33.4%, and 20.3%, respectively. These results demonstrate that CoStitch preserves the overall target structure and color appearance while achieving closer agreement with the reference in stitch texture.

4.3 Ablation Study

As summarized in Table 2, Config.A optimizes all candidate blocks without attribute-sensitive localization; Configs.B and C use only the image-state response r_1 or textual-condition response r_2, respectively; Config.D removes the color design reconstruction loss $\mathcal{L}_d$; and Config.E removes Stitch texture enhancement. As shown in Figure 8, Config.A exhibits color deviations and less distinct stitch patterns, while Configs.B and C show color or local texture mismatches when using a single response. Config.D reduces color consistency, and Config.E weakens stitch texture continuity. In comparison, the full model better preserves target structure while maintaining color appearance and fine-grained stitch characteristics. Quantitatively, the full model achieves the best GLCM-TD, SIFID, Palette EMD, and Palette ΔE_{00}, with a Silhouette Dice of 0.9593. Config.C achieves a

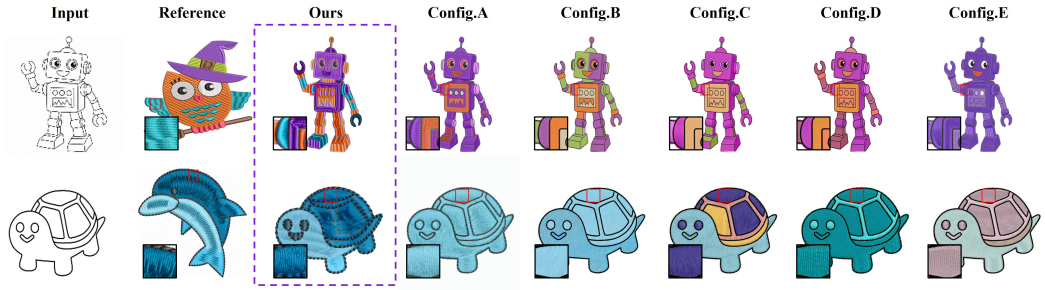


Figure 8: Qualitative results of the ablation study.

Table 2: Quantitative results of different ablation configurations.

Config.	r_1	r_2	$\mathcal{L}_d$	$\mathcal{L}_d^{x_0}$	$\mathcal{L}_e^{x_0}$	Dice↑	HFRD↓	GLCM-TD↓	SIFID↓	EMD↓	ΔE_{00} ↓
A	–	–	✓	✓	✓	0.9705	0.0060	0.1185	22.2466	39.1293	12.2967
B	✓	×	✓	✓	✓	0.9763	0.0039	0.1711	24.6908	36.1454	12.3295
C	×	✓	✓	✓	✓	0.9742	**0.0037**	0.1561	23.7646	40.0631	12.5208
D	✓	✓	×	✓	✓	**0.9794**	0.0043	0.1398	23.0299	40.6760	13.3269
E	✓	✓	✓	×	×	0.9739	0.0039	0.1770	26.2059	40.5569	12.4628
Ours	✓	✓	✓	✓	✓	0.9593	0.0048	**0.0508**	**14.1072**	**30.0398**	**10.5836**

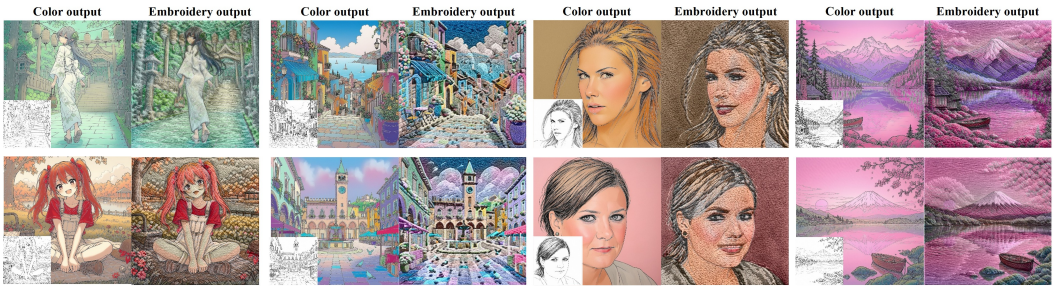


Figure 9: Additional results on complex scenes and human portraits.

slightly lower HFRD, but its higher GLCM-TD and SIFID indicate weaker overall stitch texture matching. Config.D obtains a higher Silhouette Dice because removing color design reconstruction reduces appearance variation and better preserves the target contour, but at the cost of color and texture fidelity. In comparison, the full model maintains a high level of target-structure preservation while achieving substantially better color appearance and stitch-texture consistency.

4.4 Extended Experiments

Figure 9 further demonstrates the applicability of CoStitch to complex scenes and human portraits, preserving major layouts and contours while reproducing reference color style and stitch texture.

5 Conclusion

This paper presents CoStitch for disentangled color style and stitch texture generation. By introducing the color design image as an intermediate state, CoStitch distinguishes color-related variation from stitch-texture variation, providing differentiated supervision for the two attributes. Attribute-sensitive block localization defines attribute-specific learning subsets for Color LoRA and Stitch LoRA, while successive reconstruction progressively learns color style and fine-grained stitch texture. Texture enhancement further refines local stitch characteristics while preserving the learned color style. Experiments demonstrate target-structure preservation, independent reuse and cross-reference combination of color style and stitch texture, together with applicability to complex scenes and portraits.

AI USE STATEMENT

Generative AI tools were used only for language polishing. All AI-assisted revisions were reviewed and verified by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This work focuses on embroidery image generation and does not involve human subjects, private personal data, or sensitive attributes. The embroidery images used for dataset construction were collected and processed solely for research purposes. We do not identify significant ethical risks associated with the proposed method, and all experiments were conducted in accordance with standard research integrity practices.

REPRODUCIBILITY STATEMENT

We provide the model architecture, training procedure, optimization objectives, evaluation metrics, and experimental settings in the main paper. Additional implementation details, data preparation procedures, and supplementary results are provided in the supplementary material to facilitate reproducibility.

REFERENCES

Dar-Yen Chen, Hamish Tennent, and Ching-Wen Hsu. ArtAdapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8619–8628, 2024.

X. Chen, M. McCool, A. Kitamoto, et al. Embroidery modeling and rendering. In *Proceedings of Graphics Interface 2012*, pp. 131–139. Canadian Information Processing Society, 2012.

Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. 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*, pp. 8795–8805, 2024.

Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, and Changsheng Xu. Arbitrary style transfer via multi-adaptation network. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2719–2727, 2020.

Yingying Deng, Fan Tang, Xingjia Pan, Weiming Dong, Chongyang Ma, and Changsheng Xu. StyTr^2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11326–11336, 2022.

Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, 1945.

Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using B-LoRA. In *European Conference on Computer Vision*, pp. 181–198. Springer, 2024.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023.

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

X. Guan, L. Luo, H. Li, et al. Automatic embroidery texture synthesis for garment design and online display. *The Visual Computer*, 37(9–11):2553–2565, 2021.

Robert M. Haralick, K. Shanmugam, and Its'hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, 1973.

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

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

Kibeom Hong, Seogkyu Jeon, Junsoo Lee, Namhyuk Ahn, Kunhee Kim, Pilhyeon Lee, Daesik Kim, and Hyeran Kim. AesPA-Net: Aesthetic pattern-aware style transfer networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22758–22767, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

X. Hu, C. Yang, F. Fang, et al. MSEmbGAN: Multi-stitch embroidery synthesis via region-aware texture generation. *IEEE Transactions on Visualization and Computer Graphics*, 31(9):5334–5347, 2025.

Siyu Huang, Jie An, Donglai Wei, Jiebo Luo, and Hanspeter Pfister. QuantArt: Quantizing image style transfer towards high visual fidelity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5947–5956, 2023.

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

Maxwell Jones, Sheng-Yu Wang, Nupur Kumari, David Bau, and Jun-Yan Zhu. Customizing text-to-image models with a single image pair. *ACM Transactions on Graphics*, 43(6):1–13, 2024.

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

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

Chang Liu, Viraj Shah, Anika Cui, Skanda Koppula, Jiayuan Deng, Xin Wang, et al. UnZipLoRA: Separating content and style from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23907–23917, 2025a.

H. Liu, L. Wang, Y. Zhang, et al. SaMam: Style-aware state space model for arbitrary image style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. AdaAttN: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6649–6658, 2021.

J. Ma, Q. He, G. He, et al. One-shot embroidery customization via contrastive LoRA modulation. *ACM Transactions on Graphics*, 44(6):272:1–272:15, 2025.

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

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

Litu Rout, Yujia Chen, Nataniel Ruiz, et al. RB-Modulation: Training-free stylization using reference-based modulation. In *International Conference on Learning Representations*, 2025.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. ZipLoRA: Any subject in any style by effectively merging LoRAs. In *European Conference on Computer Vision*, pp. 422–438. Springer, 2024.

Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. SinGAN: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4570–4580, 2019.

Gaurav Sharma, Wencheng Wu, and Edul N. Dalal. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30, 2005.

Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatar-Net: Multi-scale zero-shot style transfer by feature decoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8242–8250, 2018.

D. Sofronova and R. A. Angelova. A comparative analysis of the capabilities of digital embroidery software products. In *Proceedings of the XXIV Scientific Conference with International Participation FPEPM 2019*, pp. 364–369, Sozopol, Bulgaria, 2019.

Y. Takahashi and T. Fukusato. Stitch: An interactive design system for hand-sewn embroidery. In *Proceedings of ACM SIGGRAPH 2018 Posters*, pp. 1–2. ACM, 2018.

Hao Tang, Songhua Liu, Tianwei Lin, Siyu Huang, Fu Li, Dongliang He, and Xinchao Wang. MASTER: Meta style transformer for controllable zero-shot and few-shot artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18329–18338, 2023.

Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. InstantStyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024.

Linfeng Wen, Chengying Gao, and Changqing Zou. CAP-VSTNet: Content affinity preserved versatile style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18300–18309, 2023.

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

Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1395–1403, 2015.

Xin Xie, Yi Li, Hua Huang, Haiyan Liu, Daniele Panozzo, and Olga Sorkine-Hornung. Artistic style discovery with independent components. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19870–19879, 2022.

H. Yan, H. Zhang, and Z. Zhang. Learning to disentangle the colors, textures, and shapes of fashion items: A unified framework. *IEEE Transactions on Multimedia*, 26:5615–5629, 2024.

K. Yang, J. Zhou, Z. Sun, et al. Image-based irregular needling embroidery rendering. In *Proceedings of the 5th International Symposium on Visual Information Communication and Interaction*, pp. 87–94. ACM, 2012.

K. Yang, Z. Sun, C. Ma, et al. Paint with stitches: A random-needle embroidery rendering method. In *Proceedings of the 33rd Computer Graphics International*, pp. 9–12. ACM, 2016.

Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8035–8045, 2022.

Yabin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. A unified arbitrary style transfer framework via adaptive contrastive learning. *ACM Transactions on Graphics*, 42(5):1–16, 2023.

Shuke Zheng, Peng Gao, Pan Zhou, and Jing Qin. Puff-Net: Efficient style transfer with pure content and style feature fusion network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Y. Zhou, X. Gao, Z. Chen, et al. Attention distillation: A unified approach to visual characteristics transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18270–18280, 2025.

Ming Zhu, Xuanhua He, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. All-to-key attention for arbitrary style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23109–23119, 2023.

A APPENDIX

A.1 MODEL AND TRAINING SETTINGS

CoStitch is built on the pretrained SDXL model, with all backbone parameters frozen during training. Color LoRA and Stitch LoRA are inserted into the candidate attention blocks, while their trainable parameter subsets vary across reconstruction steps. All input images are resized to 1024×1024. The LoRA rank is set to 32, and the learning rate is set to 1×10^{-4}. Training is performed for a total of 800 iterations.

The first 400 iterations successively reconstruct the sketch, color design image, and embroidery image to progressively learn color appearance and stitch texture. Each reconstruction is followed by a parameter update. During sketch reconstruction, all Color LoRA parameters are updated, while Stitch LoRA is inactive. During color design reconstruction, Color LoRA parameters in the color-sensitive blocks and all Stitch LoRA parameters are updated. During embroidery reconstruction, Color LoRA participates in the forward pass without being updated, and only Stitch LoRA parameters in the texture-sensitive blocks are updated. The remaining 400 iterations are used for texture enhancement. Sketch reconstruction is stopped, Color LoRA is disabled, and only Stitch LoRA is optimized. The color design and embroidery images are successively reconstructed using the predicted x_0. The color design reconstruction updates all Stitch LoRA parameters, followed by the embroidery reconstruction, which updates Stitch LoRA parameters in the texture-sensitive blocks. This stage further refines stitch characteristics, including stitch direction, length, density, and local arrangement.

The textual conditions correspond to the visual attributes represented by different image states. Specifically, c_d denotes the textual condition of the color design image and primarily provides semantic guidance related to color appearance, while c_e denotes the textual condition of the embroidery image and further describes stitch-texture characteristics in addition to color information. During texture enhancement, c_d is used for color design image reconstruction. For embroidery reconstruction, c_e is injected only into the texture-sensitive blocks, while the remaining candidate blocks retain c_d. This conditional assignment concentrates stitch-related textual guidance on texture-sensitive blocks while preserving the color representation established by the color design image.

A.2 ATTRIBUTE-SENSITIVE BLOCK CONFIGURATION

We consider 11 attention blocks in the SDXL U-Net as candidate blocks, covering the downsampling, middle, and upsampling components. Following the attribute-sensitive block localization described in the main paper, the candidate blocks are ranked separately for color style and stitch texture, and the top four blocks are selected for each attribute. Table 3 lists the network locations of all candidate blocks and the resulting configurations.

For color learning, the selected block set is $\Theta_{\text{col}} = \{\theta_1, \theta_3, \theta_7, \theta_{11}\}$. For stitch texture learning, the selected block set is $\Theta_{\text{emb}} = \{\theta_4, \theta_7, \theta_8, \theta_{11}\}$. The two sets are determined independently and may overlap; in our configuration, θ_7 and θ_{11} are shared by both sets. These selected blocks define the restricted parameter subsets used in the corresponding reconstruction steps. Other reconstruction steps update all parameters of the active LoRA, as detailed in Appendix A.1.

A.3 INFERENCE DETAILS

After training, Color LoRA and Stitch LoRA are used independently or successively according to the generation task. The pretrained SDXL backbone remains frozen during inference, and only the corresponding LoRA parameters are activated. For an unseen target pattern, CoStitch supports color style generation, stitch texture transfer, and collaborative generation.

Color Style Generation. Given a target sketch I_s^*, we extract its Canny edges and use them as the ControlNet condition to constrain the target structure. Color LoRA is activated in the color-sensitive blocks Θ_{col}, while Stitch LoRA remains disabled. Starting from random noise, the color design image is generated as

$$I_d^* = \mathcal{G}_{\text{col}}\left(I_s^*, c_d^*; \Theta_{\text{col}}\right), \tag{8}$$

Table 3: Candidate attention blocks in the SDXL U-Net and the attribute-sensitive blocks selected for color style and stitch texture.

Block	SDXL U-Net Attention Block	Color	Texture
θ_1	`down_blocks.1.attentions.0`	✓	–
θ_2	`down_blocks.1.attentions.1`	–	–
θ_3	`down_blocks.2.attentions.0`	✓	–
θ_4	`down_blocks.2.attentions.1`	–	✓
θ_5	`mid_block.attentions.0`	–	–
θ_6	`up_blocks.0.attentions.0`	–	–
θ_7	`up_blocks.0.attentions.1`	✓	✓
θ_8	`up_blocks.0.attentions.2`	–	✓
θ_9	`up_blocks.1.attentions.0`	–	–
θ_{10}	`up_blocks.1.attentions.1`	–	–
θ_{11}	`up_blocks.1.attentions.2`	✓	✓

where $\mathcal{G}_{\text{col}}$ denotes the diffusion generation process with Color LoRA. The resulting I_d^* preserves the target structure while reproducing the primary color appearance and color relationships learned from the reference.

Stitch Texture Transfer. Given a target color design image I_d^*, we adopt Img2Img inference to generate the corresponding embroidery image. The edges extracted from I_d^* are additionally used as the ControlNet condition to preserve structural layout. During this process, Color LoRA is disabled and Stitch LoRA is activated in the texture-sensitive blocks Θ_{emb}. The inference process is written as

$$I_e^* = \mathcal{G}_{\text{emb}}\left(I_d^*, c_e^*; \Theta_{\text{emb}}\right), \tag{9}$$

where $\mathcal{G}_{\text{emb}}$ denotes the Img2Img diffusion process with Stitch LoRA. The input color design image provides the target structure and color appearance, while Stitch LoRA introduces the learned stitch direction, length, density, and local arrangement.

Collaborative Generation. For collaborative generation, the two inference processes are performed successively. Given a target sketch I_s^*, Color LoRA first generates an intermediate color design image I_d^*. This intermediate result is then used as the input to Stitch LoRA to generate the final embroidery image I_e^*:

$$I_e^* = \mathcal{G}_{\text{emb}}\left(\mathcal{G}_{\text{col}}\left(I_s^*, c_d^*; \Theta_{\text{col}}\right), c_e^*; \Theta_{\text{emb}}\right). \tag{10}$$

Color LoRA and Stitch LoRA can be obtained from different reference samples, enabling the color style of one reference and the stitch texture of another to be combined on the same target structure.

B Additional Experimental Results

B.1 Additional Qualitative Comparisons

We provide additional qualitative comparisons to further evaluate CoStitch under a broader range of reference and target configurations. In addition to the comparisons presented in the main paper, we examine three complementary settings: general reference-target pairs, multiple target patterns under the same embroidery reference, and embroidery references with diverse stitch appearances. These settings allow us to further examine the preservation of target structure, the consistency of reference appearance transfer across different targets, and the adaptability to variations in stitch texture.

Figure 10 presents additional comparisons on a broader set of reference-target pairs. The reference embroideries and target designs vary in pattern structure and color composition. Across these examples, CoStitch preserves the contours and spatial organization of the target designs while reproducing the primary color appearance and fine-grained stitch characteristics of the references. In contrast, baseline methods exhibit limitations across samples. Some results introduce structural characteristics from the reference, while others show color deviations or weakened and less continuous stitch patterns. These results are consistent with the main paper and further demonstrate CoStitch's ability to transfer embroidery appearance across diverse reference-target pairs.

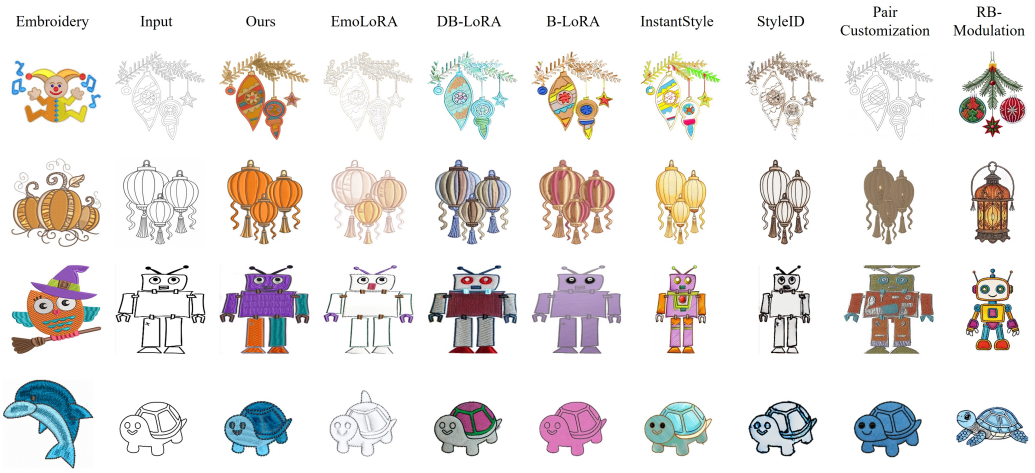


Figure 10: Additional qualitative comparisons with existing methods across different reference-target pairs.

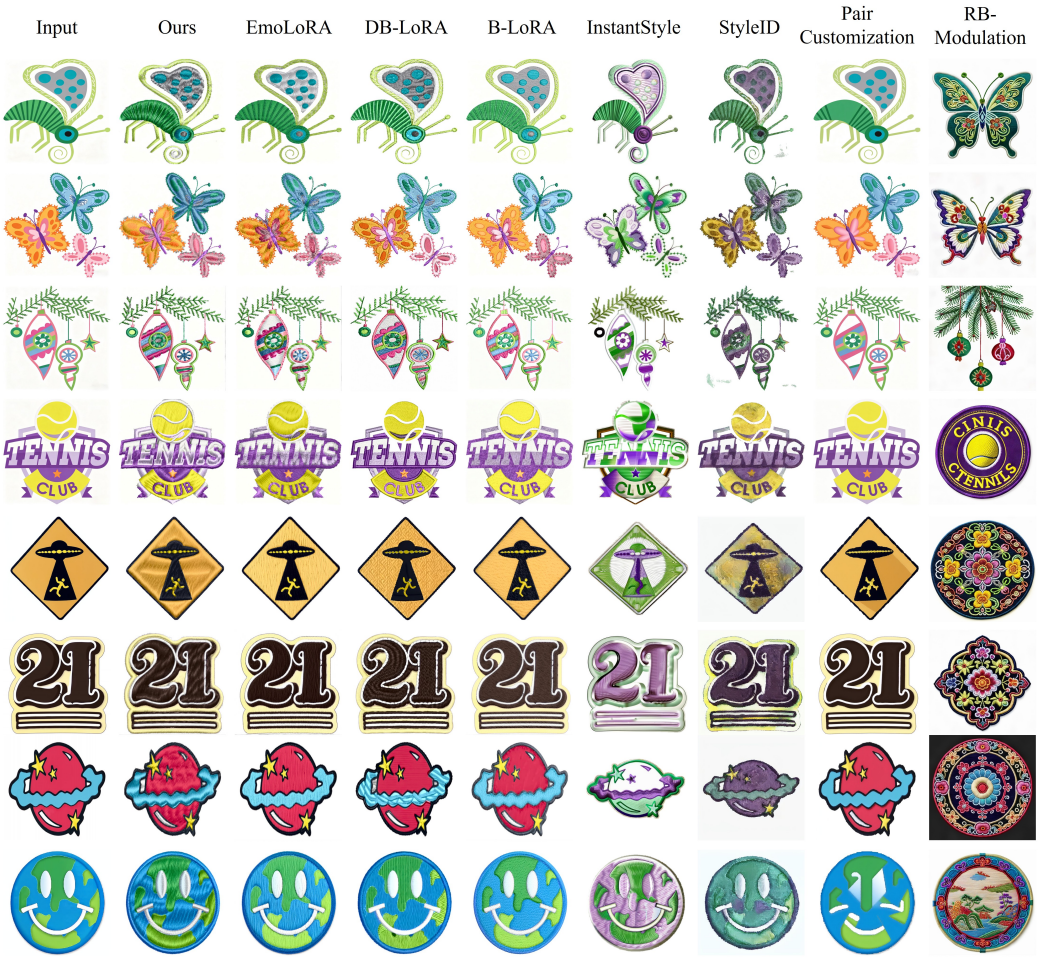


Figure 11: Qualitative comparisons using the same embroidery reference across different target patterns.

Figure 11 evaluates appearance-transfer consistency by fixing the embroidery reference while varying the target pattern. CoStitch adapts the reference appearance to different target structures while preserving their contours and spatial layouts. In contrast, the baselines exhibit varying degrees of reference-structure leakage, color variation, or reduced stitch-texture consistency. These results further demonstrate that the learned embroidery appearance can be reused across different target structures.

Figure 12 further compares stitch texture transfer using references with diverse stitch appearances. CoStitch produces reference-dependent stitch patterns while preserving the target structure and established color appearance, with local stitches adapting to target contours. In contrast, several baselines show weakened stitch structures, less distinctive textures, or appearance deviations from the references. Together with the fixed-reference comparison in Figure 11, these results show that CoStitch can consistently reuse the same embroidery appearance across different target structures while adapting stitch texture to different references.

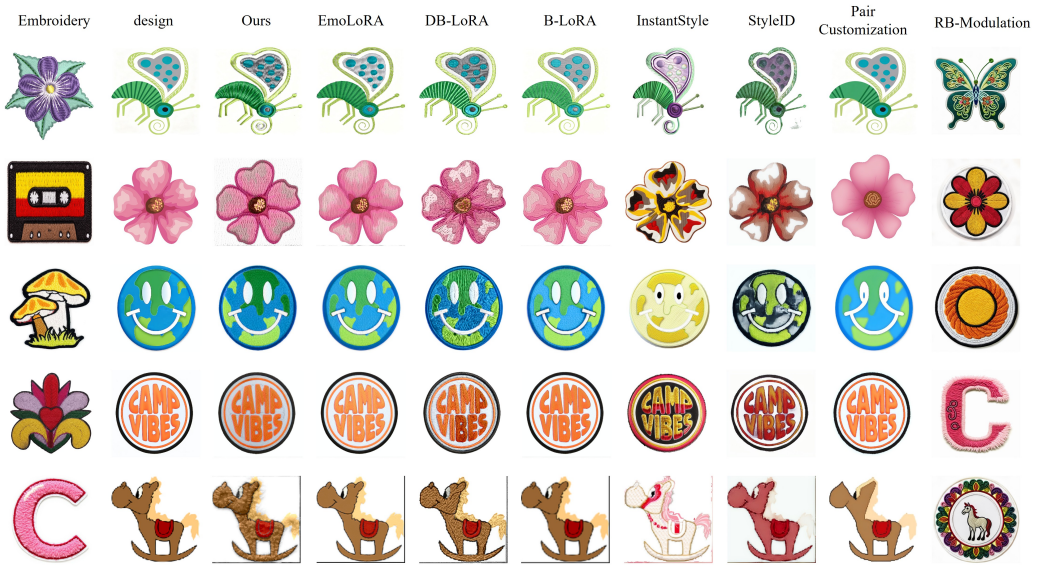


Figure 12: Qualitative comparisons of stitch texture transfer using embroidery references with diverse stitch appearances.

B.2 Additional Color Style Generation Results

Figure 13 presents additional color style generation results. In each row, a single color reference is applied to multiple target sketches with different structures. CoStitch preserves the target contours and spatial layouts while reproducing the dominant colors and color relationships of the corresponding reference. Across different target patterns, the generated results maintain consistent color characteristics while adapting to each target structure. These results further show that the learned color representation can be reused without transferring the structure of the reference image.

B.3 Additional Stitch Texture Transfer Results

Figure 14 presents additional stitch texture transfer results. In each row, a single embroidery reference is applied to multiple target color designs with different structures. CoStitch preserves the target structure and color appearance while transferring fine-grained stitch characteristics, including direction, length, density, and local arrangement. Across different target patterns, the generated results maintain consistent stitch characteristics while adapting their local organization to the geometry of each target design. The generated stitch patterns therefore adapt to the local geometry of each target rather than reproducing the spatial structure of the reference embroidery.



Figure 13: Additional color style generation results. Each row uses one color reference for multiple target patterns.



Figure 14: Additional stitch texture transfer results. Each row uses one embroidery reference for multiple target patterns.