Supplementary - 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

In this supplementary material, we include more details, results, and discussions about the experiments. Specifically, we first provide the implementation details of our method in Sec. 1. Then we expand the evaluation on embroidery customization in Sec. 2, including discussion on metrics, more qualitative comparisons, ablation studies, and computational cost. Subsequently, we explore the potential of our customized embroideries in transforming real-world workflows in Sec. 3, in regard to the usage for preview and presale, fabrication acceleration, as well as other virtual applications. To verify the capability of our method in decoupling style and content, we conduct experiments on three more visual attribute transfer tasks in Sec. 4, including photo-to-artwork in Sec. 4.1, sketch-to-color in Sec. 4.2, and appearance transfer in Sec. 4.3. Finally, we provide the prompts used for embroidery generation in Sec. 5.

ACM Reference Format:
Jun Ma, Qian He, Gaofeng He, Huang Chen, Chen Liu, Xiaogang Jin, and Huamin Wang. 2025. Supplementary - One-shot Embroidery Customization via Contrastive LoRA Modulation. *ACM Trans. Graph.* 44, 6, Article 271 (December 2025), 21 pages. https://doi.org/10.1145/3763290

1 IMPLEMENTATION DETAILS

We implement our method with SDXL V1.0 and generate images in 1024×1024. The rank of LoRA in our experiments is set to 32. We adopt Adam as the optimizer and set the learning rate to 1e-4. For contrastive learning, we adjust the weight for each loss term instead of tuning the learning rate, simply for convenience. We set the weight for complementary data as 0.1 to avoid deterioration in embroidery style from noisy generated data. The weight for contrastive loss is set to 0.001 due to its fast convergence. We train 400 steps for the first stage and 200 steps for the second stage.

2 EVALUATION ON EMBROIDERY CUSTOMIZATION

In this section, we delve deeper into our evaluation on embroidery customization. Importantly, we first discuss the strengths and limitations of the metrics, which underscores the necessity of conducting user studies to complement the quantitative evaluation. Then, we present qualitative comparisons on additional reference embroidery images, followed by ablation studies on block selection, color correction, and multi-style training. Finally, we report the computational cost of our method and compare it with existing approaches.

2.1 Metrics

HRDF. We compare the compliance of the generated image's embroidery style to the reference with their High-Frequency Ratio Difference (HFRD). To mitigate the impact of foreground size variations, we segment each image into 16 patches of 256×256 pixels. Subsequently, we discard any patches where over 50% of the pixels are near white in color. For the remaining patches, we first convert them to grayscale images and then apply FFT (Fast Fourier Transform) to obtain their spectral representation. To distinguish between low-frequency and high-frequency components, we utilize a central circular mask. The high-frequency region is defined as the portion of the spectrum where the distance from its center exceeds 0.3 times the edge length. We compute the average high-frequency energy ratio across all patches and report its absolute difference from the corresponding value of the reference.

LPIPS. We use LPIPS to assess the preservation of design content. For each generated embroidery image, we first conduct Gaussian blur with kernel size 15 to alleviate the influence of embroidery texture, and then compute its LPIPS with the input design image.

Discussion. We utilize HFRD and LPIPS to assess image-based customization outcomes. However, given the subtle variations and intricate structures inherent to embroidery features, HFRD occasionally fails to provide an accurate reflection of the objective. This can be observed in Fig. 1, where one positive example and one negative example are presented. Notably, generated embroideries that are quite dissimilar to the reference may still exhibit very low HFRD values. Moreover, when comparing two embroidery generations in terms of their design content preservation using LPIPS, the one with better color compliance might receive a worse score due to the

*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

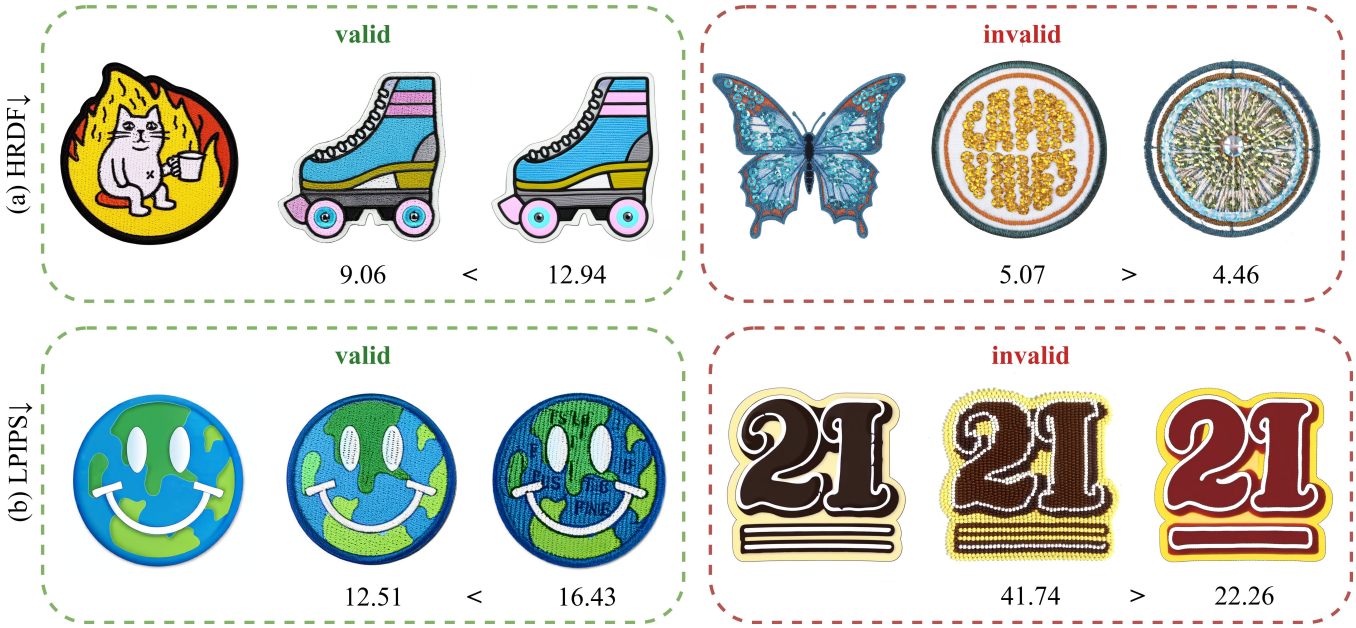

Fig. 1. Analysis on HRDF and LPIPS. For each metric, we show one positive example that align with the visual quality (valid), and one negative example where the visually superior generation yields a worse quantitative score (invalid).

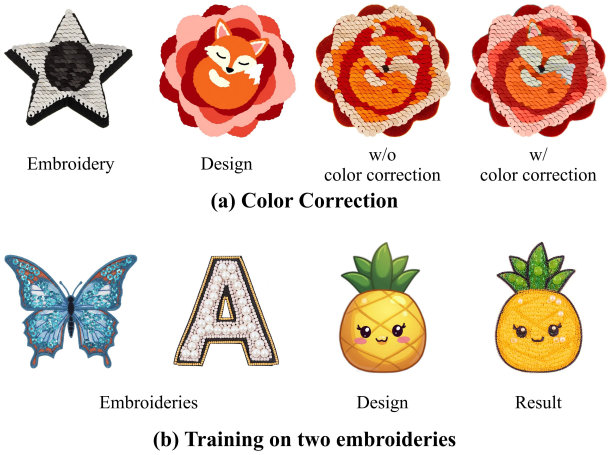

Fig. 2. Ablation on color correction and multi-style training. (a) Color correction on sequins leads to noticeable misalignment (Row 1, Column 4). (b) Training a single EmoLoRA on two embroidery styles results in bead deformation and unintended fusion of pearls and sequins.

influence of embroidery features. Due to the noise in these metrics, the quantitative results fail to distinguish between our method, DB-LoRA, and B-LoRA. Therefore, we rely on more direct assessments through user studies.

2.2 More Qualitative Comparison

In Fig. 3, we show visual comparisons on more reference embroideries, to further illustrate the generalization capability of our method.

Moreover, we conduct qualitative comparisons with Attention Distillation [Zhou et al. 2025] in both image-based and text-based generation, and with Analogist [Gu et al. 2024] in image-based generation, as shown in Fig. 4. Attention Distillation leverages attention features from pretrained diffusion models without explicit disentanglement for fine-grained styles. As a result, it tends to treat overall appearance—including color and texture—as transferable style, but still fails to preserve clear structures for beads, pearls, and sequins. In contrast, Analogist fails to apply embroidery-like textures and does not maintain the structural consistency of the input design.

2.3 Ablation Study on Color Correction

We conduct ablation studies to examine the effect of the color correction module. For flat and chenille embroideries, where boundary alignment is particularly important, all methods employing ControlNet during inference (Ours, DB-LoRA [Ruiz et al. 2023], B-LoRA [Frenkel et al. 2025], and InstantStyle [Wang et al. 2024]) are equipped with color correction. To mitigate the color bias introduced by ControlNet, we convert the generated images into LAB space and substitute their AB channels with those of the corresponding design images. For sequin and bead embroideries, however, the generation process inherently alters the internal structure, and direct AB-channel replacement would cause unnatural artifacts (Fig. 2 (a)), so color correction is not applied. We present quantitative results with and without color correction in Tab. 1. Without color correction, our method still achieves stronger color fidelity. Note that

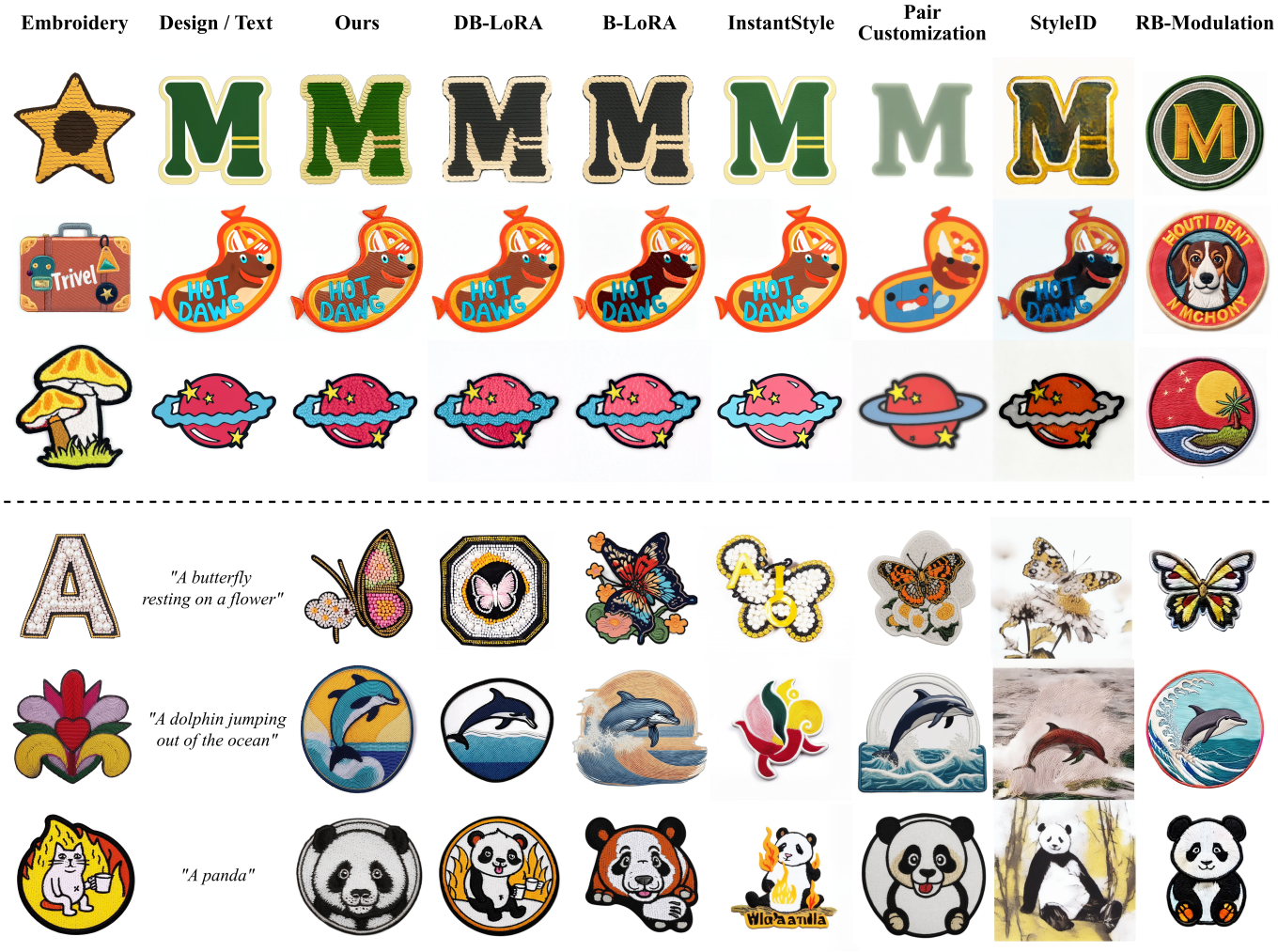

Fig. 3. More comparisons on one-shot embroidery customization. (Please zoom-in to see the detailed textures.)

Table 1. Ablation on color correction. The best results are highlighted in bold, and the second-best are underlined. We employ Histogram-Loss and LPIPS to quantitatively evaluate the preservation of design color and content, respectively.

Metric	Ours	DB-LoRA	B-LoRA	InstantStyle	PairCustomization	StyleID	RB-Modulation
Histogram-Loss↓ (w/o)	<u>44.00</u>	45.18	46.57	46.11	**43.99**	45.75	48.87
Histogram-Loss↓ (w/)	**26.59**	<u>28.62</u>	30.57	32.23	43.99	45.75	48.87
LPIPS↓ (w/o)	<u>18.73</u>	19.40	21.10	**13.87**	22.14	21.96	65.18
LPIPS↓ (w/)	<u>14.37</u>	14.54	14.92	**7.72**	22.14	21.96	65.18

PairCustomization and InstantStyle preserves the design content with minimal embroidery features.

2.4 Ablation Study on Multi-Style Training

Distinct styles like embroidery and color are often encoded by different model weights. Focusing on one style may limit the model's ability to represent others. Moreover, high intra-class variation within embroidery styles can hinder convergence and lead to style fusion. For instance, as shown in Fig. 2 (b), training a single EmoLoRA on two embroidery styles results in bead deformation and an unintended fusion of pearls and sequins. This observation highlights the challenge of multi-style training and suggests that separate models or explicit disentanglement configurations may be necessary to faithfully preserve distinct style patterns.

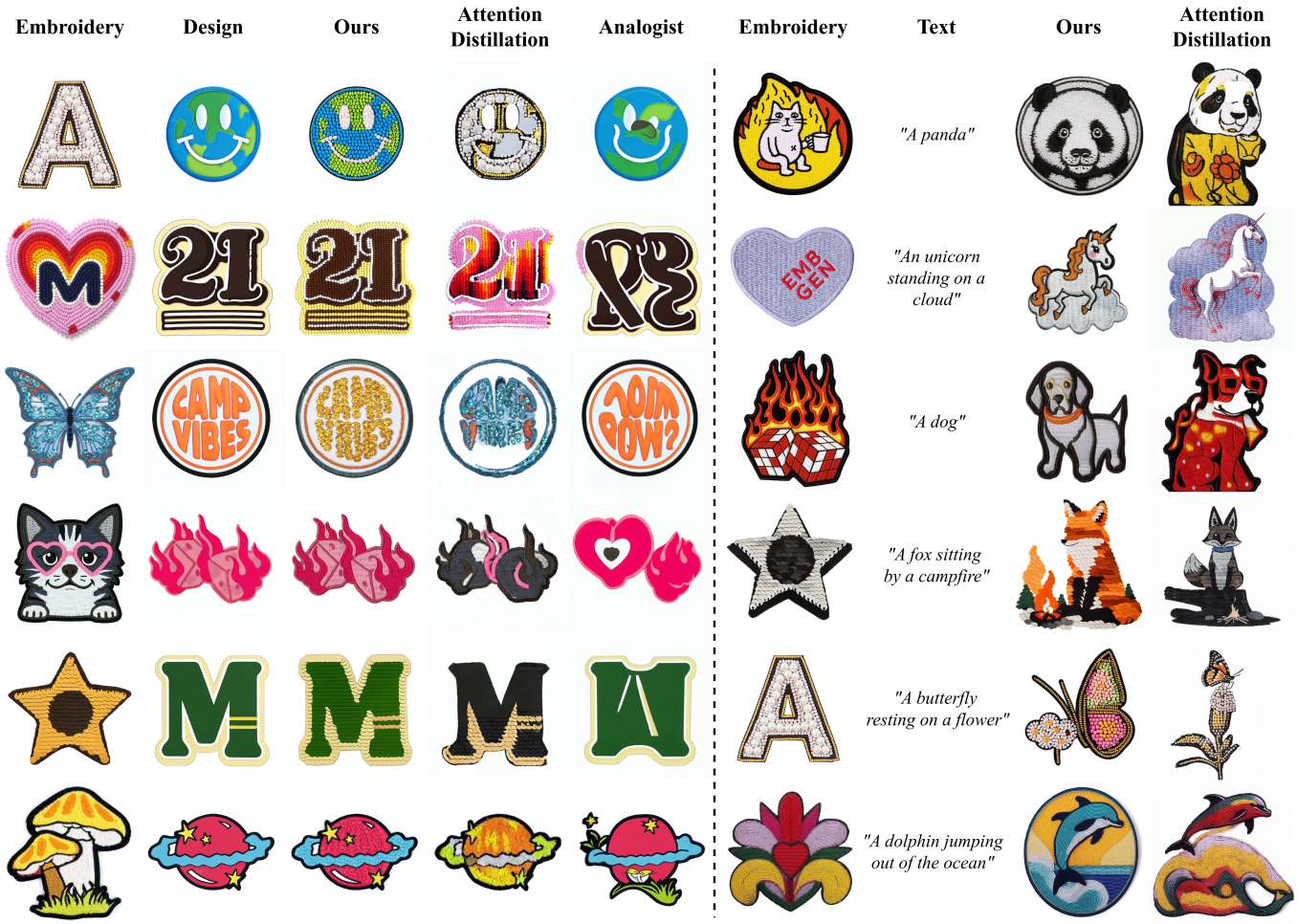

Fig. 4. Qualitative comparison with Attention Distillation [Zhou et al. 2025] in both image-based and text-based generation, and with Analogist [Gu et al. 2024] in image-based generation.

Table 2. Comparison on computational cost.

Metric	Ours	DB-LoRA	B-LoRA	InstantStyle	PairCustomization	StyleID	RB-Modulation	Attention Distillation
Training-VRAM (GB)	20.96	**9.78**	12.11	-	11.33	-	-	-
Training-Time (s)	1220	**106**	310	-	294	-	-	-
Inference-VRAM (GB)	13.97	14.00	13.94	16.76	14.94	18.28	19.10	**3.65**
Inference-Time (s)	7.63	8.32	**7.16**	11.25	44.22	16.35	65.93	16.62

2.5 Ablation Study on Block Selection

We conduct more ablation studies on our block selection. From all blocks labeled 1-11 in the SDXL base model, we choose *Block-2,3,7,8* as our final model. To verify if this is the best choice, we conduct experiments on four reference embroideries, as shown in Fig. 5. Firstly, we try four individual blocks with the lowest average cosine similarity. *Block-2* and *Block-3* are from down_blocks, demonstrating minimal structural features. *Block-7* and *Block-8* are from up_blocks and capture different structures, yet still barely perceptible. Using two blocks, *Blcoks-2,8*, *Blocks-2,3* or *Blocks-7,8*, captures more structural features than using a single block, while none of them encompass complete sequins or pearls. Selecting blocks with higher average cosine similarity, *Blocks-4,5*, demonstrates minimal embroidery style. Using down_blocks only, *Blocks-1,2,3,4*, shows no structural features, while using up_blocks only, *Blocks-7,8,9,10*, misses important details. Combining down_blocks and up_blocks but with higher average cosine similarity, *Blocks-1,4,6,9*, indeed fails

Fig. 5. Ablation study on block selection. Pink flower design (Row 2, Column 1) © Vecteezy.

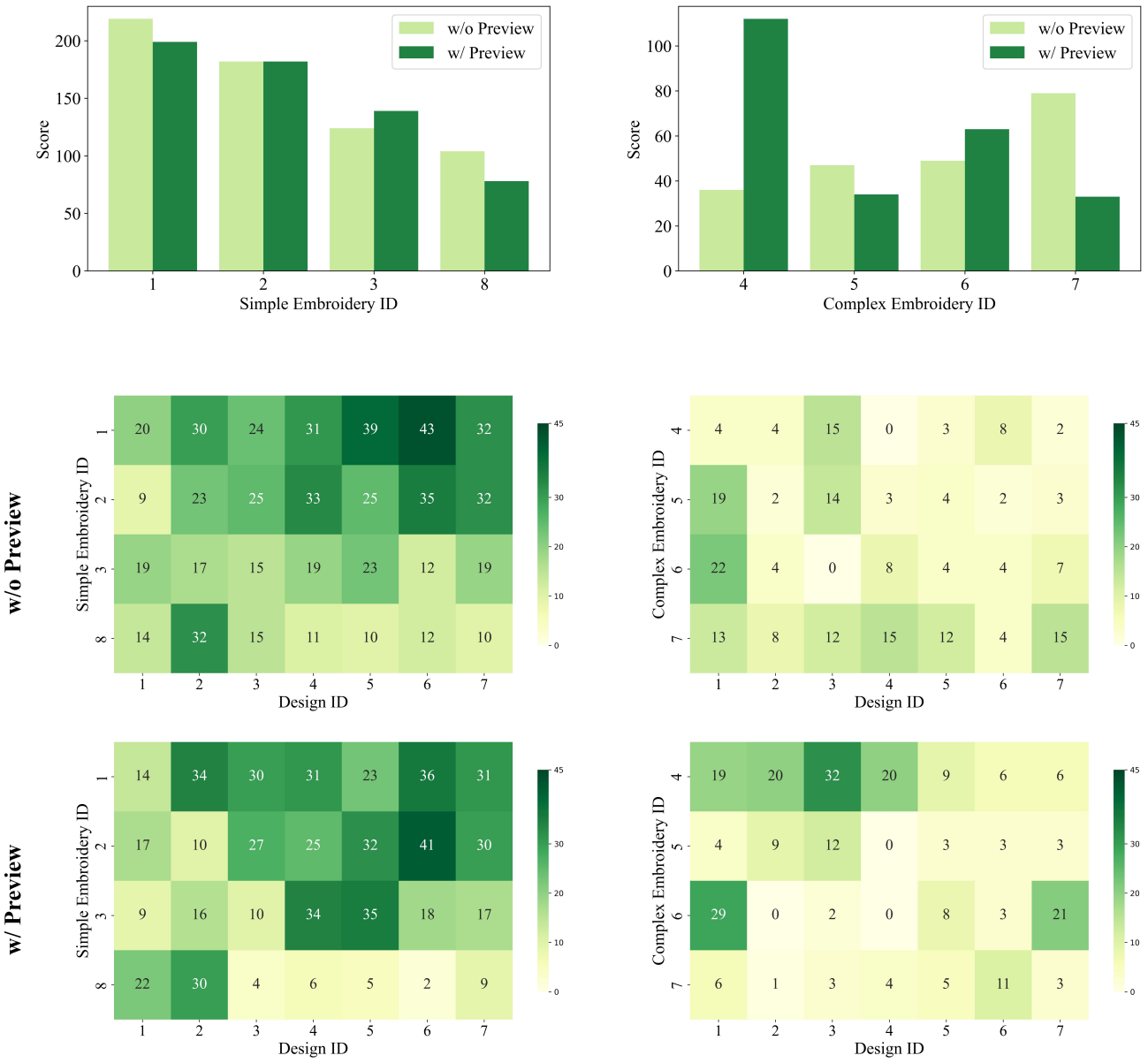

Fig. 6. Statistics of user preference.

to capture embroidery structures. Using more blocks than ours while missing one of our blocks, *Blocks-1,2,4,5,7,8* or *Blocks-2,3,4,5,8,10*, also misses important structural features. Using more blocks including all of our blocks also has worse integration with input design content due to entanglement with the reference content. *Blocks-2,3,4,5,7,8* has worse results than *All Blocks* because it squeezes more structure features into the semantic blocks (two blocks out of six in total) and thus has more severe entanglement issues. To summarize, using the four selected blocks with the lowest average cosine similarity in our method achieves the best embroidery customization results.

2.6 Computational Cost

We compare the training and inference costs of different methods on an NVIDIA RTX 4090 GPU, as reported in Table 2. While our method requires extended training to learn fine-grained styles, it enables efficient inference without relying on inversion or optimization during test time.

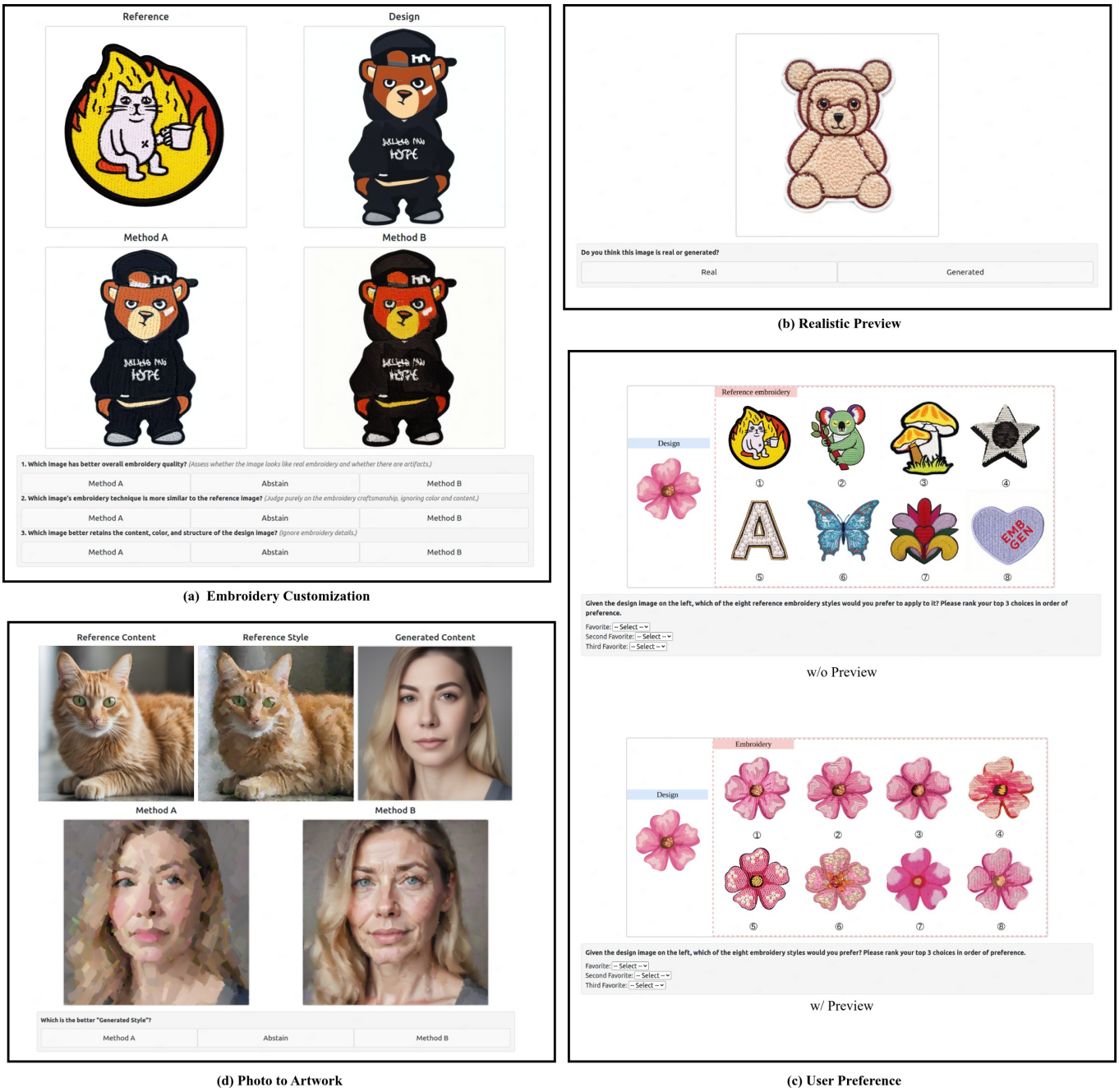

Fig. 7. User Study Interfaces. Pink flower design © Vecteezy.

3 TRANSFORMATION TO EMBROIDERY WORKFLOWS

In this section, we investigate the potential of the customized embroideries in transforming real-world embroidery workflows. To achieve this, we first conduct user studies on applications in preview and presale, bridging the communication between producers and customers. Then we explore the usage of the customized embroideries in fabrication with the Wilcom EmbroideryStudio. Finally, we discuss other potential of our method in generating high-quality embroidery and design images.

3.1 Preview and Presale

To investigate whether embroidery customization can facilitate presale, thereby improving the alignment between production and

Fig. 8. Reference embroidery style, input design image, our generated embroidery and the corresponding digitized embroidery result. The digitizing process is much more straight-forward with our generated embroidery than with the input design image. Pink flower design © Vecteezy.

sales and alleviating inventory pressure, we conduct several user studies.

Firstly, we evaluate whether the generated embroideries are sufficiently realistic. To achieve this, we generate 50 high-quality embroideries with our method in any style with either image or text inputs, and collect 50 real embroideries. Then we build a questionnaire for each user with 15 real and 15 generated random sampled from the prepared set. The interface is as shown in Fig. 7 (b), we ask the user to choose whether they think the image is real or generated.

Additionally, we choose eight reference embroideries in different styles and seven design images to generate customized embroideries for preview and presale, with the results shown in Fig. 9. Then, we design a second questionnaire as illustrated in Fig. 7 (c), where users are first presented with an interface displaying a design image alongside eight reference embroidery styles and asked to select their top three preferences in order. Subsequently, the customized results are shown, and users are asked to make their selections again. Each user answers to seven pairs of *w/o Preview* and *w/ Preview* questions.

We collect answers from 20 users for the these two questionnaires, all with no specialized knowledge of embroidery. Before answering questions, we show them 10 real embroidery patches in different styles, and then ask them to answer these two questionnaires in order. Note that the questions in each questionnaire are in random order and these two questionnaires share no images in common.

In the first questionnaire, 67.3% of the 300 votes for our generated images mistakenly identified them as real embroideries, while 19.7% of the 300 votes for real embroideries were misidentified as generated. These results suggest two key insights: (1) the generation quality of our method is convincingly realistic; and (2) users were making informed judgments rather than random guesses.

For the second questionnaire, we assign scores to the results: 3 points for the most preferred, 2 for the second, and 1 for the third. The final score statistics are summarized in Fig. 6. The Embroidery IDs and Design IDs are aligned with the ones in Fig. 9 and Fig. 7. We present the results by categorizing the embroideries into simple and complex sets, and compare the scores between *w/o Preview* and *w/ Preview*. The heatmaps display the total scores of customized results for each combination of reference embroidery and design. The histograms show the total score summed across all designs for each reference embroidery.

From the results in the histograms, the relative ranking of the four embroidery styles in the simple group remains consistent, suggesting that the presence or absence of a preview makes little difference. In contrast, for the complex embroidery group, the relative preferences shifted significantly: the style that received the lowest score before previewing became the highest-rated after the preview was provided. This result can significantly influence producers' decisions on which style to prioritize for production.

This phenomenon is similarly evident in the heatmaps. We analyze the data from both the customer and producer perspectives. From the customer perspective, we calculated the average proportion of changes across all participants in their top-1, 2, 3 selections: **72.1%**, **58.2%**, and **47.1%**. For example, 58.2% means that when given the preview, on average, users changed more than half of their top-2 selections. The results illustrate the importance of preview in customization scenarios. From the producer's perspective, we conducted separate analyses for simple and complex embroideries. Assuming we need to select the top-3, 5, or 10 embroidery results for mass production, our goal is to align these selections as closely as possible with consumer preferences to maximize potential sales. To achieve this, we refer to the heatmaps to identify the top-3, 5, 10 candidates. During this process, we observed that the proportions of changes in selection for simple embroideries were **33.3%**, **60.0%**, and **30.0%** respectively, while for complex embroideries, the corresponding change rates were **66.7%**, **60.0%**, and **70.0%**. These results highlight the value of using customized embroidery previews, particularly in scenarios involving mass production, where aligning with consumer preferences is crucial.

3.2 Fabrication

Based on the preview and presale statistics, we simulate the digitization and production process for top-ranked embroidery styles. We select Embroidery-6-Design-1 in Fig. 9, as it exemplifies a range of stitching styles and ornamental components. We employ Wilcom EmbroideryStudio to simulate the digitizing process, and the results are shown in Fig. 8.

With our generated embroidery images, manufacturable files can be produced through manual tracing. The process is as follows: first, the image is imported into Wilcom software. Then, the dominant color tones are extracted from the image and matched to actual embroidery thread color codes to ensure accurate color reproduction during digitization. The image is then resized to the target embroidery dimensions (e.g., 100mm × 100mm) to facilitate real-world production. Next, a layer-by-layer analysis is conducted, allowing

Fig. 9. Our embroidery customization results for image-based generation. Pink flower design © Vecteezy.

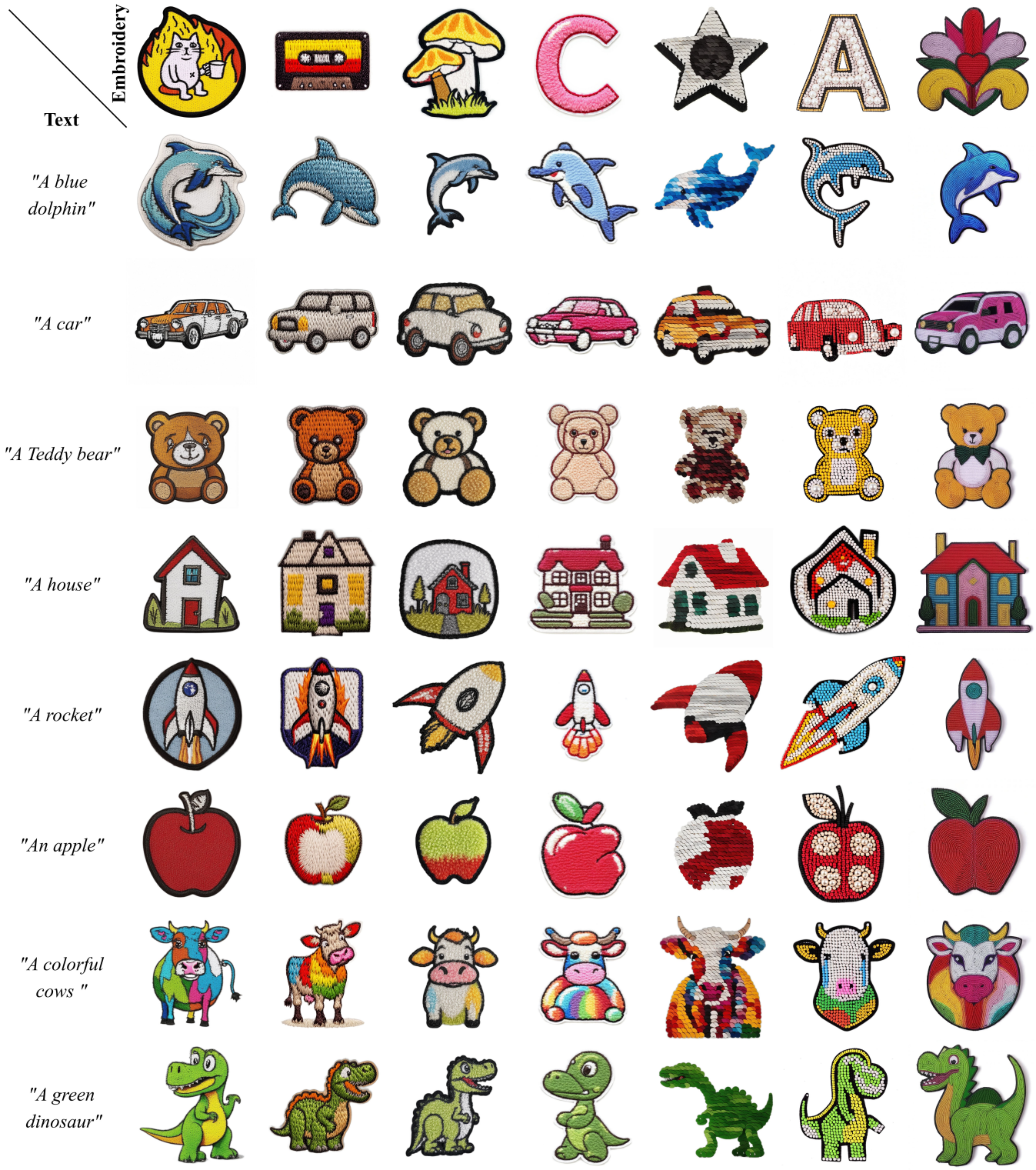

Fig. 10. Our embroidery customization results for text-based generation.

(a) Embroidery to Design

(b) Virtual Display

Fig. 11. Embroidery-to-design results and virtual display.

digitizers to manually apply stitch types to each region—for example, using Tatami stitches for base fill and Satin stitches for edge reinforcement. After that, decorative elements such as sequins and beads are analyzed by measuring their size and color in the image, and corresponding parameters (e.g., sequin diameter, shape, and color code) are set in the software to match the real materials. Finally, the decorations are accurately placed according to the reference image.

Our generated embroidery images offer the following benefits for fabrication: First, they can be traced to create manufacturable files, offering a rapid initialization for subsequent manual refinement and thereby accelerating the digitization process; second, they support the preview of diverse design outcomes through random generation, reducing iterative cycles of digitization and confirmation; third, our customization can serve as a realistic rendering module, enhancing the realism of digitized embroidery and facilitating communication between producers and customers.

3.3 Other Applications

Embroidery Data Generation. We show more customization results using our method in both image-based and text-based settings. In Fig. 9 and Fig. 10, we demonstrate the capability of our method in generating high-quality embroidery data, which can alleviate challenges posed by data scarcity in this domain.

Design Image Recovery. Our embroidery-to-design module is able to recover well-aligned design images based on reference embroideries, as shown in Fig. 11 (a). This feature enables extracting design content from an existing embroidery, facilitating new style synthesis for this design.

Virtual Display. Our generated embroidery can be overlaid onto garments, bags, hats, and other items to provide more intuitive and effective visual previews, through integration with ACE++ [Mao et al. 2025]. The examples are shown in Fig. 11 (b). Given our generated embroidery images, ACE++ takes either image or text input for the target object, and produces the decoration result. For image-based input, the embroidery can be placed at a specified location; for text-based input, it is automatically positioned on the generated object based on the textual description. This application further assists designers and customers in refining their choices.

4 GENERALIZATION TO OTHER STYLES

We evaluate our method on diverse styles to demonstrate its effectiveness in separating style and content. To achieve this, we conduct comparisons with prior art on three tasks: artistic style transfer, sketch colorization, and appearance transfer.

4.1 Photo to Artwork

We first compare to PairCustomization [Jones et al. 2024], which also learns artistic style from a single image pair. On this task, we achieve comparable results with PairCustomization on in-domain style transfer, while slightly better performance for cross-domain generalization, which demonstrates the efficacy of our method in style disentanglement. We provide the details as below.

4.1.1 Adaptation and Implementation. We follow PairCustomization and use the same pair-wise data construction strategy for fair comparisons, incorporating both artist-created artworks and images generated using external stylization methods.

We collect five training pairs, each representing a distinct content category (e.g., cat, dog, woman, man, landscape) and an associated visual style (e.g., painted, digital art, posterization, painting, cartoon), as shown in Fig. 12 and Fig. 13. Specifically, two content-style pairs ("dog" and "man") are provided by PairCustomization. The remaining three pairs are constructed using SDXL-generated photo content and stylized with three different techniques, including White-box Cartoonization [Wang and Yu 2020], Posterization (filter-based), and Stylized Neural Painting [Zou et al. 2021].

Given a photo-artwork pair, we then apply the similarity metric by analyzing the SDXL model to identify the blocks that are most correlated to the target artistic style. After that, we apply the same two-stage contrastive LoRA learning strategy (*EmoLoRA*) as described in our embroidery pipeline. For block selection, we choose 3–4 blocks with notably low average cosine similarity to balance the trade-off between style-content disentanglement and effective style learning. Our training pairs and the corresponding block selection results are shown in Fig. 12 and 13, where the selected blocks are highlighted in red. Using these blocks, we fine-tune the model with

Table 3. Quantitative comparisons on photo to artwork. Best and second-best results are highlighted in bold and underlined. SDXL results are shown for reference only and excluded from ranking.

Metric	Ours	DB-LoRA	B-LoRA	PairCustomization	*SDXL*
Style (All)↓	**0.158**	0.220	<u>0.171</u>	0.183	*0.153*
Content (All)↓	<u>0.204</u>	0.274	**0.069**	0.217	*0.000*
Style (in-domain)↓	**0.156**	0.249	<u>0.181</u>	0.202	*0.156*
Content (in-domain)↓	<u>0.214</u>	0.331	**0.091**	0.254	*0.000*
Style (cross-domain)↓	**0.160**	0.191	<u>0.162</u>	0.163	*0.149*
Content (cross-domain)↓	0.193	0.217	**0.048**	<u>0.179</u>	*0.000*

paired prompts such as "a photo of a cat" (following PairCustomization) and "a photo of a cat in [emb] style", enabling it to capture both the underlying content structure and the desired stylistic attributes.

The complementary data generation is slightly different from embroidery, as obtaining style image based on content is easier than separating content from style image for this task. We generate a set of new images with different content using the base SDXL model, and then stylized using the updated model from the first stage to construct a complementary dataset, which is subsequently used for the second-stage contrastive learning.

Our inference pipeline follows the same framework as in the embroidery task. For text-based generation, we use prompts with "in [emb] style" to guide stylization. For image-based generation, we employ SDEdit[Meng et al. 2021] to add noise to the input image, and then apply our model for denoising. To conduct fair comparisons with PairCustomization, we do not leverage ControlNets to preserve content. Rather, we set the timestep to activate trained LoRA weights, as the hyperparameter for content preservation.

4.1.2 Experiments and Results. Following the experimental setup of PairCustomization, our evaluation focuses on the text-based generation setting. The objective is to assess whether the stylized outputs preserve the structure, semantics, and color characteristics of the base SDXL model outputs while successfully applying the target visual style.

Our method is compared against three baselines: DB-LoRA, B-LoRA, and PairCustomization. DB-LoRA and B-LoRA are trained solely on the style image, using the prompts "a photo of a cat in [emb] style" (trained for 400 steps) and "a [emb]" (trained for 1000 steps), respectively. PairCustomization is trained on the same paired content-style data as ours, following their original implementation and hyperparameter for both training and inference.

One key observation is that the fidelity of stylized outputs with respect to the base model is highly sensitive to the specific timestep at which LoRA is activated. Similar to PairCustomization, we adopt a controlled LoRA activation strategy during text-based generation. Specifically, during the early denoising steps ($t > T_s$), LoRA remains disabled and only the base content prompt is used. In the later steps ($t \leq T_s$), LoRA is activated alongside the stylized prompt. This progressive activation helps the model maintain core content structure while gradually applying the desired style. The activation timestep T_s is tuned for each style to achieve an optimal trade-off between stylization strength and content preservation. For fairness, the same timestep-controlled strategy is also applied to DB-LoRA and B-LoRA. Notably, PairCustomization similarly adopts timestep gating to enhance style-content alignment and further introduces an inference algorithm that preserves the original denoising path while injecting controllable style guidance.

Fig. 14 and Fig. 15 illustrate the influence of different LoRA activation timesteps for a specific style. The figures present the corresponding training data, the output from the original SDXL model, and the stylized results generated by each method under varying LoRA activation timesteps. Denoising begins at step 1000. When LoRA is activated from the very beginning (i.e., at step 1000), all methods tend to deviate substantially from the pretrained model's output, often compromising content fidelity. For each baseline, the best-performing configuration is selected through qualitative inspection based on the trade-off between content preservation and stylization strength, ensuring a fair and representative comparison.

We present the qualitative comparison results in Fig. 12, Fig. 13. For each pair, we report both in-domain and cross-domain performance (e.g., training on cats while testing on cats and humans) to evaluate the generalization capabilities of each method across different content domains.

Our method demonstrates relatively stable performance across both in-domain and cross-domain scenarios, suggesting stronger generalization due to better style-content decoupling during training.

DB-LoRA shows limited ability to disentangle style from content. Even with timestep control, it often overfits to the style image, leading to content distortion and compromised fidelity. In contrast, B-LoRA exhibits relatively weak stylization effects, indicating difficulties in capturing and transferring subtle artistic attributes. PairCustomization performs reasonably well on in-domain samples but struggles to generalize across categories. For example, when trained on the *woman-posterization* style, it occasionally fails to apply the style consistently to cross-category samples such as dogs—particularly in stylizing salient foreground regions—and also degrades on landscape scenes (e.g., Fig. 14 and Fig. 13).

In comparison, our method achieves comparable results to PairCustomization on in-domain tasks and exhibits more stable behavior in cross-domain conditions. These results demonstrate the effectiveness of our contrastive learning framework in promoting consistent style-content disentanglement, as well as enhancing generalization across content categories and visual styles.

We adopt the same metric, DreamSim [Fu et al. 2023], as PairCustomization to evaluate style consistency with GT and content

Table 4. Quantitative comparisons on sketch to color.

Metric	Ours	DB-LoRA	B-LoRA	ColorizeDiffusionV1.5	ColorizeDiffusionV2	InstantStyle	PairCustomization
F1-Score↑	<u>67.58</u>	54.86	63.55	65.54	**68.99**	63.42	60.79
Histogram-Loss↓	43.06	<u>40.10</u>	52.60	46.98	**37.71**	48.26	56.17

consistency with SDXL generation. As shown in Tab. 3, our method achieves the best style consistency with relatively low content drift in both in-domain and cross-domain settings, demonstrating its effectiveness in disentangling style and content across diverse visual styles and content categories. Note that these metrics are not fully aligned with the actual objectives: SDXL scores highest on both, while B-LoRA largely preserves content, both with minimal style yet still attain competitive style distance scores.

4.1.3 User Study. We also conducted a user study to quantitatively evaluate three baseline methods against our proposed approach. To assess generalization, both in-domain and cross-domain testing were performed, following the protocol established by PairCustomization. For each training pair, we generated 20 in-domain samples and 20 cross-domain samples (five from each of the remaining categories), forming a test dataset designed to evaluate both style fidelity and content preservation under varying contexts.

The study followed the same pair-wise comparison protocol as our embroidery user study. Each of the 20 participants was randomly assigned 90 sample pairs, consisting of comparisons between our method and one of the baselines. The comparison frequencies were balanced across all methods to ensure fairness.

As illustrated in Fig. 7 (d), participants were shown a reference content-style image pair along with a newly generated content image in the first row. They were then presented with two stylized outputs—each generated by a different method—and asked to choose the one they found more appropriate. This interface is simliart to the one used in PairCustomization.

The results show that the probabilities of DB-LoRA, B-LoRA, and PairCustomization being preferred over our method were **29.32%**, **7.13%**, and **40.04%**, respectively. This further supports that our method achieves a better balance between content preservation and effective stylization, and demonstrates stronger disentanglement between content and style representations compared to the baselines.

4.2 Sketch to Color

We also apply our method to sketch colorization, to verify our capability in decoupling style (color and shading) with content (semantics and layouts) in this domain. The results suggest that our method effectively captures the style from a training pair and blends it compatibly with new content.

4.2.1 Adaptation and Implementation. For pair-wise data construction, we mimic the content image from sketch by extracting a Canny edge map from the reference color image. In this way, the content is roughly defined as semantics and layouts, while the style is the separated color and shading. We collect four color images, including two landscape scenes and two anime characters as in Fig. 17, and then obtain four color-Canny training pairs. For the similarity metric, we observe that these pairs exhibit a similar distribution of average cosine similarity across different attention blocks, suggesting a consistent style-content relationship. Therefore, we adopt the average among these four pairs as in Fig. 16 (left) and select the four blocks highlighted in red.

Then we apply our contrastive LoRA learning strategy. For an example of Canny-color image pair, the corresponding prompts are "an anime girl" and "an anime girl in [emb] style", respectively. After completing the first-stage training, we generate complementary data in stylized outputs using diverse content prompts. Canny edge maps are then extracted from these outputs and paired with their corresponding stylized images to form an augmented dataset. This dataset is subsequently used in the second training stage to improve content-style disentanglement through contrastive learning.

During inference, we employ ControlNet-Canny to preserve the content of the input sketch. The generation is guided by prompts such as "a landscape in [emb] style" or "an anime girl in [emb] style", depending on the target content, for text-based synthesis. This setup enables the model to perform sketch colorization while maintaining accurate structural alignment with the original sketch.

4.2.2 Experiments and Results. Evaluation is conducted under both in-domain (e.g., training and testing on similar content such as anime sketches) and cross-domain (e.g., training on landscapes while testing on anime sketches) settings.

We compare our method against several baselines: DB-LoRA and B-LoRA, which are trained solely on reference style images using the prompts "an anime girl in [emb] style" (400 steps) and "a [emb]" (1000 steps), respectively; InstantStyle [Wang et al. 2024], which directly inputs the reference image for style transfer with an IP-Adapter; and PairCustomization, which is trained on the same Canny–color image pairs and evaluated using their ControlNet-Canny inference pipeline. For compatibility with the setting and claims in PairCustomization, the style prompt "[emb]" is replaced with "colorful" in their training. All baselines are evaluated on the same test sketches using consistent ControlNet settings. Additionally, we include two state-of-the-art sketch colorization models—ColorizeDiffusionV1.5 [Yan et al. 2025b] and ColorizeDiffusionV2 [Yan et al. 2025a]—with their publicly released weights and inference pipelines.

As shown in Fig. 17, qualitative comparisons highlight key differences across methods. Our approach successfully handles the sketch colorization task by effectively transferring both color and shading styles. It achieves a strong balance between structural preservation and stylistic consistency across both in-domain and cross-domain settings.

In contrast, DB-LoRA, although capable of capturing the reference color style, often suffers from content leakage. For instance, in the fifth row of Fig. 17, the short hairstyle in the sketch is incorrectly

Table 5. Quantitative comparisons on appearance transfer.

Metric	Ours (a)	Ours (b)	DB-LoRA	Attention Distillation	Cross-Image Attention	InstantStyle
Struct.-Preservation↑	<u>0.937</u>	0.902	0.788	0.872	0.825	**0.943**
Appearance-Fidelity↓	1.220	<u>1.096</u>	1.545	**1.001**	1.101	1.311

replaced with the long hairstyle from the reference image, indicating that the model has overfitted to both the style and content of the reference.

B-LoRA, InstantStyle, and PairCustomization frequently show color drifts and fail to maintain coherent stylistic features. ColorizeDiffusionV1.5 and V2 exhibit reasonably good performance on in-domain anime characters, probably due to their large-scale training on animation-style datasets. However, their generalization to cross-domain scenarios—such as transferring landscape colors to character sketches or even landscape-to-landscape stylization—remains limited, often producing visually unnatural results.

These observations confirm that our method not only captures the style characteristics of color and shading effectively, but also benefits from style-content disentanglement, allowing it to generalize well to new sketches. This results in stylized outputs that preserve the input semantic structure while faithfully reflecting the reference style, even under cross-domain conditions.

We further evaluate structural similarity using the F1-Score between the Canny edges of the generated image and the input sketch, and color consistency using Histogram-Loss with the reference image, shown in Tab. 4. Although ColorizeDiffusionV2 achieves the best quantitative scores, it is trained on 6.5 million images, whereas our method is trained on a single image at test time. Importantly, our approach produces more natural colorization in cross-domain scenarios (e.g., landscapes to portraits), demonstrating a better trade-off between structure preservation and style fidelity. DB-LoRA, despite achieving lower Histogram-Loss, exhibits content leakage due to insufficient color-content disentanglement, which negatively impacts structural consistency.

4.3 Appearance Transfer

We extend our method to transfer more complex visual attributes, specifically texture, in appearance transfer, a task akin to sketch colorization but with richer styles and looser structural constraints. The results demonstrate our method's ability to capture appearance features while disentangling them from structural content using selected style blocks. When all blocks from our trained EmoLoRA are used, the structural content of the reference can also be compatibly blended with the target structure.

4.3.1 Adaptation and Implementation. For pair-wise data construction, we adopt similar setting as in sketch colorization, and extract Canny maps from reference appearance to obtain content images. Thus the content is also defined as the fine-grained structural information, containing layouts and semantics. As a result, the style is defined as the rich appearance features aside from the content.

We construct four appearance-Canny training pairs, sourced from photorealistic images with rich appearance features. Then we then compute the average cosine similarity of their attention features, and observe similar patterns as sketch colorization. Two image pair examples and the averaged similarity heatmap are shown in Fig. 16 (right), and we select the same four blocks for appearance as for color.

Based on the pair-wise data and selected attention blocks, we apply our EmoLoRA with paired prompts such as "a photo of a cake" and "a photo of a cake in [emb] style". Different from previous tasks, the second-stage contrastive learning here must avoid introducing additional appearance cues from cross-category or cross-instance content, as this could alter the intended style.

Therefore, we generate image variants using the same training prompts, introducing slight perturbations to the reference image to avoid injecting extraneous appearance features. Canny edge maps are then extracted from these variants to form consistent pairs for contrastive learning.

During inference, we adopt our image-based generation pipeline with ControlNet-Canny. The content structure is obtained through extracting HED[Xie and Tu 2015] edge maps from input structure images, as HED maps preserve the rough outline structures of content images while removes fine-grained elements like fur or repetitive textures. We add noise to the input structure image with SDEdit and send the corresponding HED map to ControlNet-Canny, along with prompts like "a photo of a cake in [emb] style", to generate the final output image.

4.3.2 Experiments and Results. We compare our method with several baselines. DB-LoRA is trained solely on the reference image using the same prompt and 400 training steps. InstantStyle inputs the appearance reference directly to an IP-Adapter during inference. Both methods adopt the same ControlNet inference configuration as ours to ensure fair comparison. Attention Distillation [Zhou et al. 2025] is reproduced according to the authors' official setup, where the reference and content images are optimized jointly using a content loss weight of 0.2 over 200 steps. Cross-Image Attention [Alaluf et al. 2024] is evaluated with its default settings.

As shown in Figure 18, our method achieves more faithful appearance transfer while preserving structural consistency. Two inference settings highlight the level of style-content disentanglement of our approach: **Ours(a)** uses only the selected style blocks, and successfully transfers the reference appearance while introducing little additional structures; in contrast, **Ours(b)** activates all EmoLoRA blocks during inference, smoothly blending the reference appearance and structure into the target structure specified by the HED maps.

In the Taj Mahal example, Ours(a) accurately captures the white stone texture without distorting the target HED structure, whereas Ours(b) introduces dome-like features from the reference, resulting in a result of both appearance transfer and structure "augmentation". Similarly, in the third example featuring a car, Ours(a) transfers the

color and gloss of the car paint, while Ours(b) additionally incorporates structural elements such as the rear wing and taillights, demonstrating its ability to transfer appearance and augment structure with decoupled style and content representations.

Although DB-LoRA employs the same block configuration as Ours(b), it fails to integrate appearance and structure coherently due to the absence of disentanglement training, leading to structural artifacts. This contrast underscores the effectiveness of our contrastive learning strategy in decoupling style and content within LoRA blocks. Attention Distillation often produces incomplete shapes and unnatural blending, suggesting that it struggles to balance appearance and structure during transfer. Cross-Image Attention suffers from low visual fidelity and introduces noticeable artifacts. InstantStyle fails to capture the fine-grained appearance features of the reference and frequently yields semantically inconsistent results.

We use the same metrics as Cross-Image Attention, measuring structure preservation via IoU of foreground masks and appearance fidelity via Gram matrix differences. As shown in Tab. 5, our method achieves a balanced trade-off between structure and appearance: Ours(a) attains high structure preservation while maintaining competitive appearance fidelity, indicating effective style-content disentanglement. Ours(b) further improves appearance transfer while still preserving the structure reasonably well, demonstrating a alternative overall balance.

These comparisons demonstrate that our training framework not only successfully separates appearance from structure in the selected style blocks, but also achieves better integration with new structures using all blocks, thanks to the inherently more disentangled representations.

5 PROMPTS FOR EMBROIDERY GENERATION

In this section, we provide the prompts used for embroidery generation, in complementary data generation and text-based evaluation.

5.1 Prompts for Complementary Data Generation

For complementary data generation, we preset 10 prompts for all references:

(1) "A yellow dog in [emb] style",
(2) "A silver robot in [emb] style",
(3) "A red car in [emb] style",
(4) "A yellow train in [emb] style",
(5) "A green house in [emb] style",
(6) "A blue boat in [emb] style",
(7) "A brown horse in [emb] style",
(8) "A blue bird in [emb] style",
(9) "A green tree in [emb] style",
(10) "A pink rose in [emb] style".

5.2 Prompts for Text-based Generation

For text-based generation, we preset 20 prompts. With each prompt, we generate two random results for ours, DB-LoRA [Ryu 2022] and B-LoRA [Frenkel et al. 2025]. Since our EmoLoRA effectively decouples embroidery style from image content, including semantic layout, our text-based embroidery generation is dependent on the original text-to-image generation of SDXL base model. To make the generation more friendly for embroidery production, we add "simple background, white background" to all prompts. We list the prompts as follows:

(1) "A patch of a dog in [emb] style",
(2) "A patch of a cat in [emb] style",
(3) "A patch of a car in [emb] style",
(4) "A patch of a house in [emb] style",
(5) "A patch of a panda in [emb] style",
(6) "A patch of a fox sitting by a campfire in [emb] style",
(7) "A patch of a dragonfly hovering over a pond in [emb] style",
(8) "A patch of a rocket ship flying through stars in [emb] style",
(9) "A patch of a cactus in a colorful pot in [emb] style",
(10) "A patch of a penguin sliding on ice in [emb] style",
(11) "A patch of a sunflower blooming under a rainbow in [emb] style",
(12) "A patch of a butterfly resting on a flower in [emb] style",
(13) "A patch of a dolphin jumping out of the ocean in [emb] style",
(14) "A patch of a rabbit holding a carrot in [emb] style",
(15) "A patch of an owl perched on a moonlit branch in [emb] style",
(16) "A patch of a raccoon peeking out of a trash can in [emb] style",
(17) "A patch of a unicorn standing on a cloud in [emb] style",
(18) "A patch of a flamingo standing in a pond in [emb] style",
(19) "A patch of a hedgehog holding an apple in [emb] style",
(20) "A patch of a snail with a colorful shell in [emb] style".

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

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.

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.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344* (2023).

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.

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.

Black Forest Labs. 2024. FLUX. https://github.com/black-forest-labs/flux.

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).

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

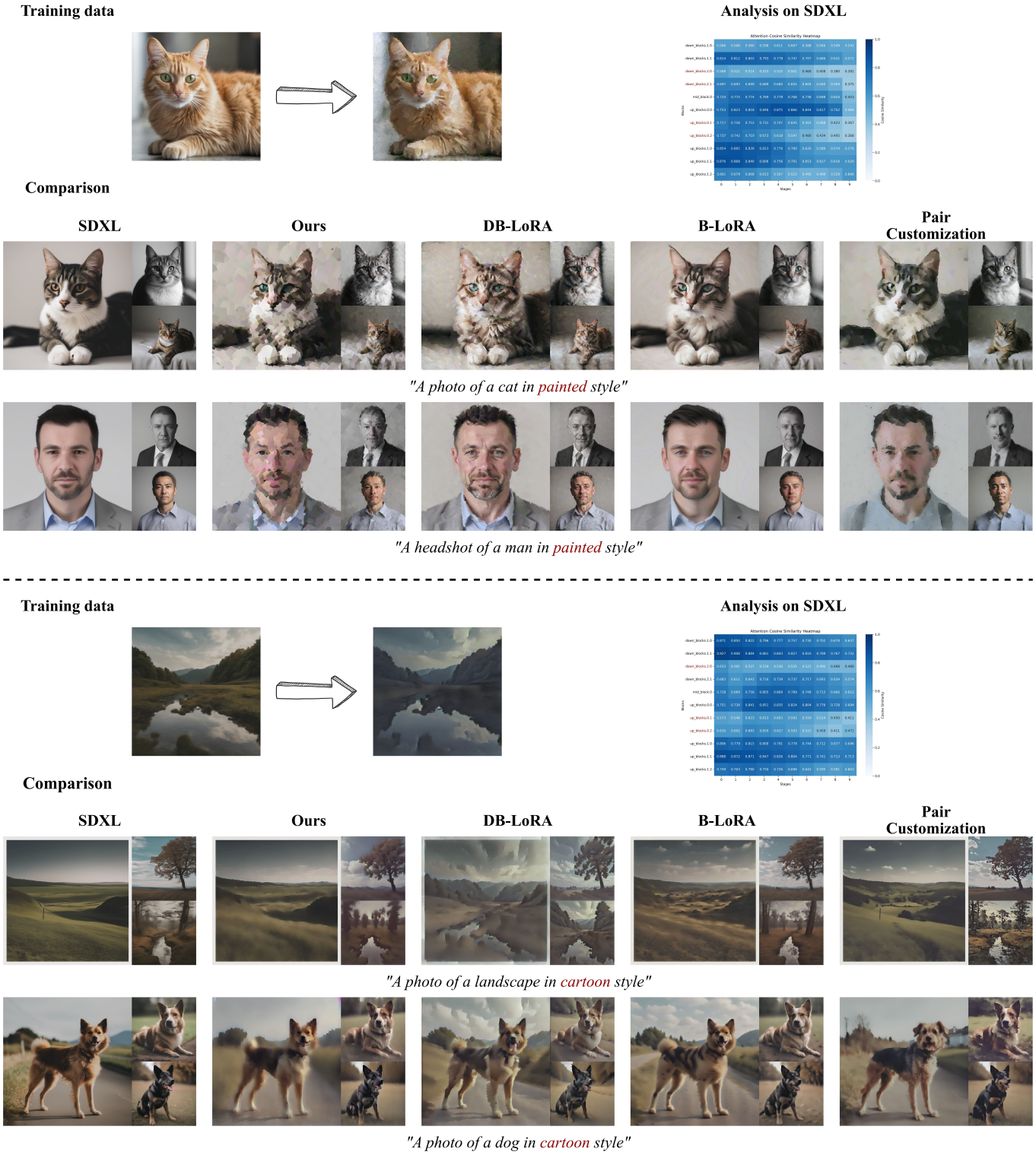

Fig. 12. Artistic style transfer results on in-domain and cross-domain cases. Top: The "Cat–Painted" training pair and corresponding feature similarity analysis on SDXL. The second row shows in-domain stylization (cat) and cross-domain generalization (man) using different methods. Bottom: The "Landscape–Cartoon" training pair and SDXL feature analysis. We present stylization results for in-domain (landscape) and cross-domain (dog) cases.

Fig. 13. Artistic style transfer results on in-domain and cross-domain cases. Top: The "Woman–Poster" training pair and SDXL feature analysis, along with stylization results for in-domain (woman) and cross-domain (landscape) examples. Bottom: The "Man–Painting" training pair and corresponding SDXL feature analysis. We show stylization results for in-domain (man) and cross-domain (cat) cases. The man training data © PairCustomization [Jones et al. 2024].

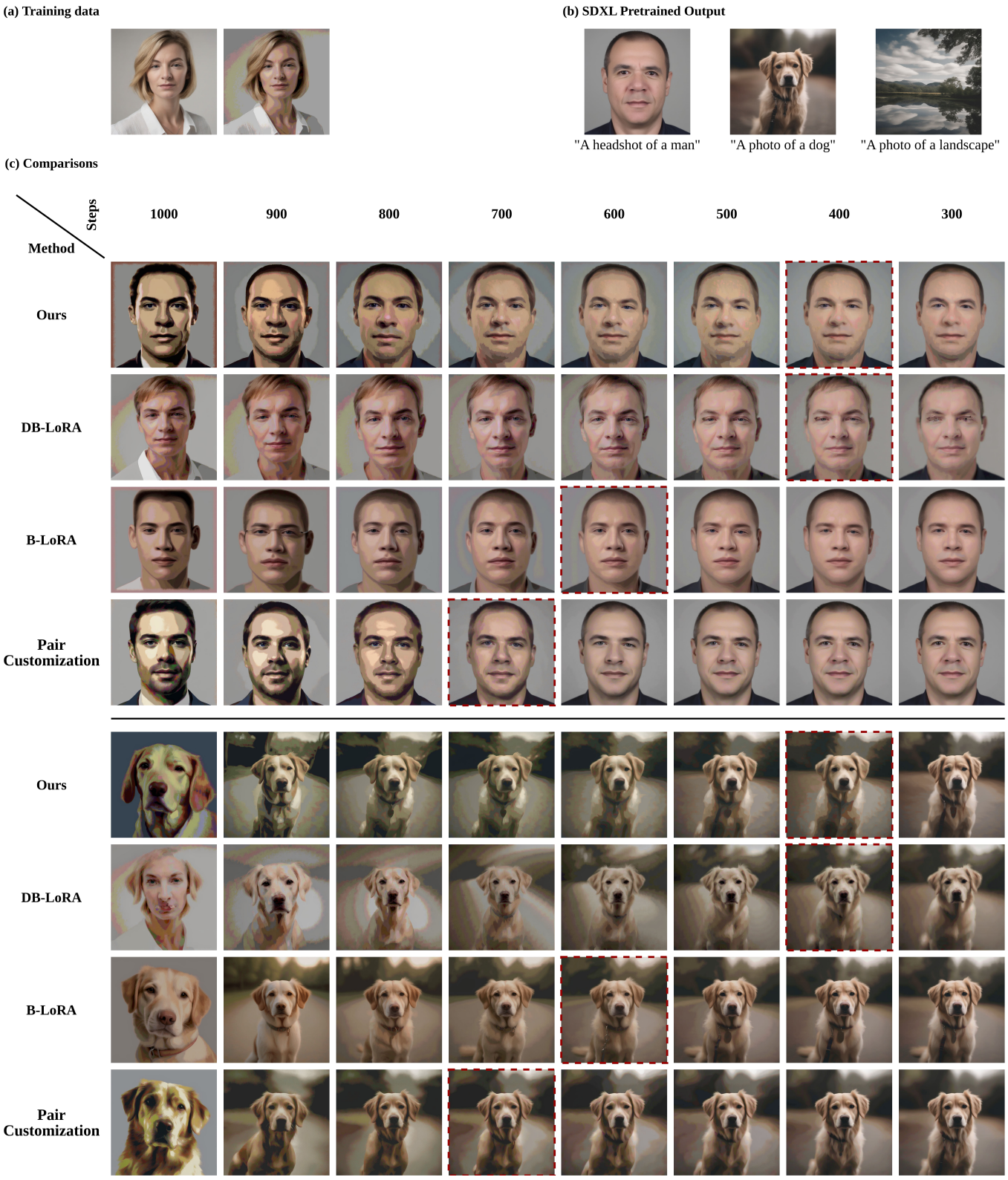

Fig. 14. Impact of LoRA activation timesteps. We show the training data, the generation from SDXL, and stylized results at different LoRA activation timesteps. The optimal timestep for each method is highlighted.

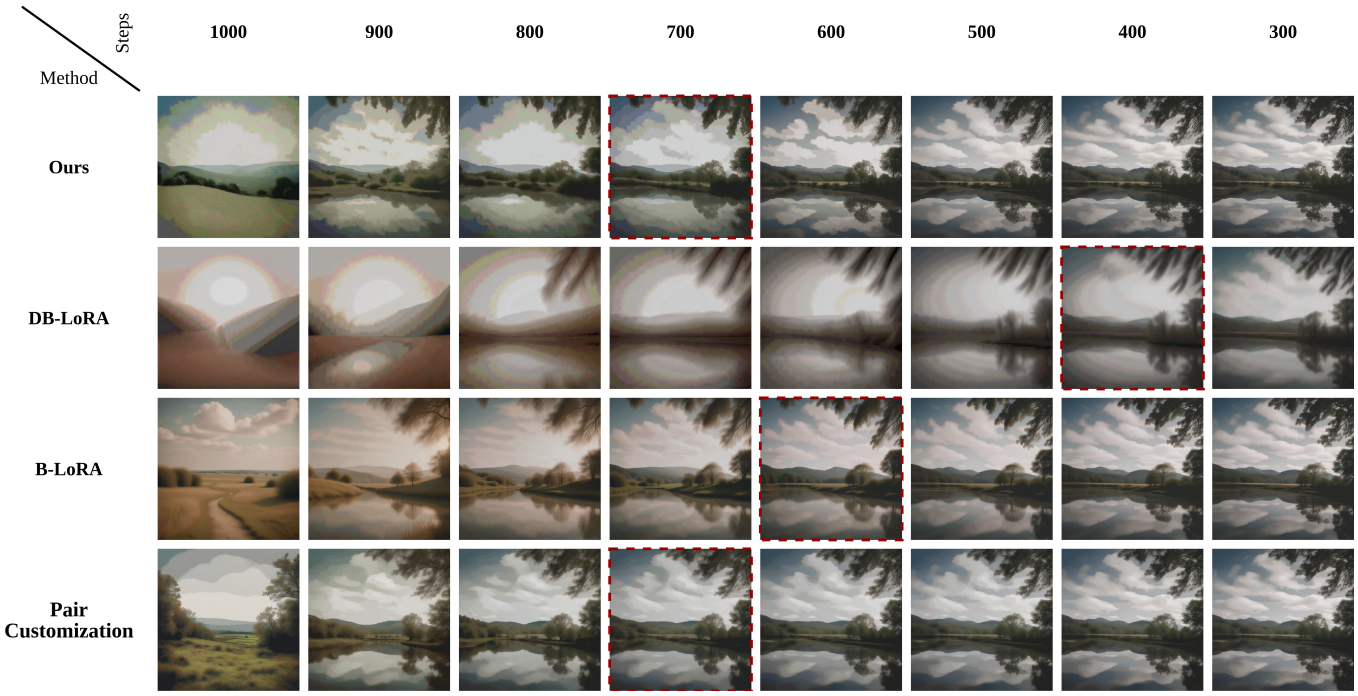

Fig. 15. Impact of LoRA activation timesteps. Comparison of stylized outputs at various LoRA activation steps for a cross-domain (landscape) example.

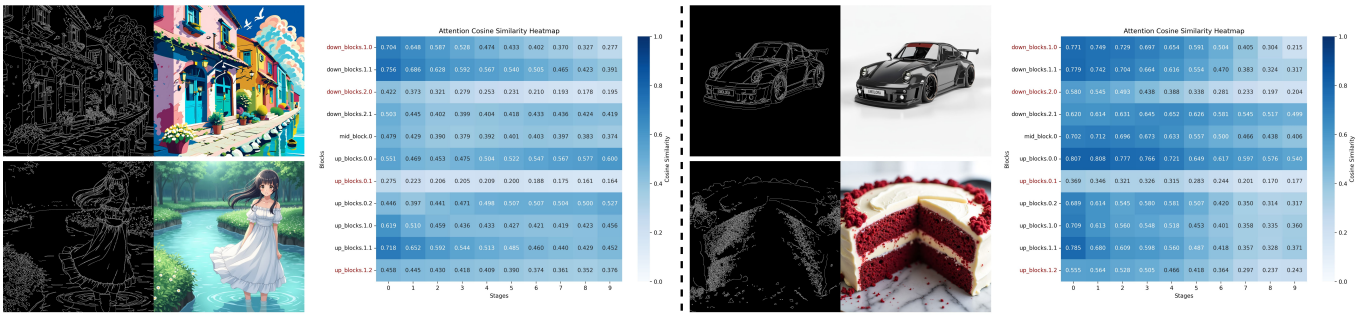

Fig. 16. Training pair examples and attention-based block selection for sketch colorization and appearance transfer. Left: Two training pair examples from the task of sketch colorization, along with the averaged similarity heatmaps. Right: Two training pair examples from the task of appearance transfer, along with the averaged similarity heatmaps.

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).

Xinrui Wang and Jinze Yu. 2020. Learning to cartoonize using white-box cartoon representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8090–8099.

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

Dingkun Yan, Xinrui Wang, Yusuke Iwasawa, Yutaka Matsuo, Suguru Saito, and Jiaxian Guo. 2025a. ColorizeDiffusion v2: Enhancing Reference-based Sketch Colorization Through Separating Utilities. *arXiv preprint arXiv:2504.06895* (2025).

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

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).

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.

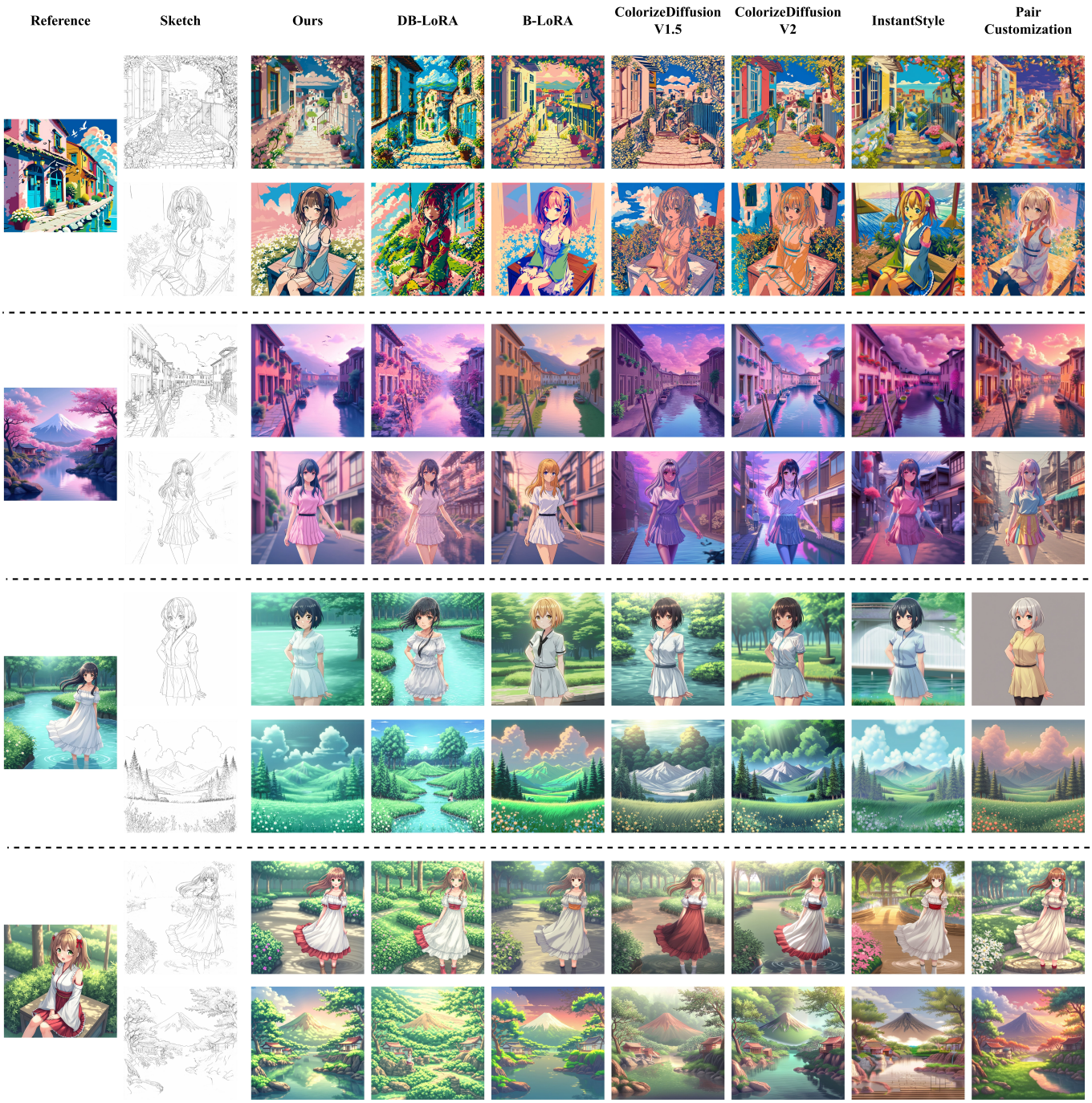

Fig. 17. Qualitative comparison on the sketch colorization task. All anime character images are generated using the FLUX [Labs 2024] model.

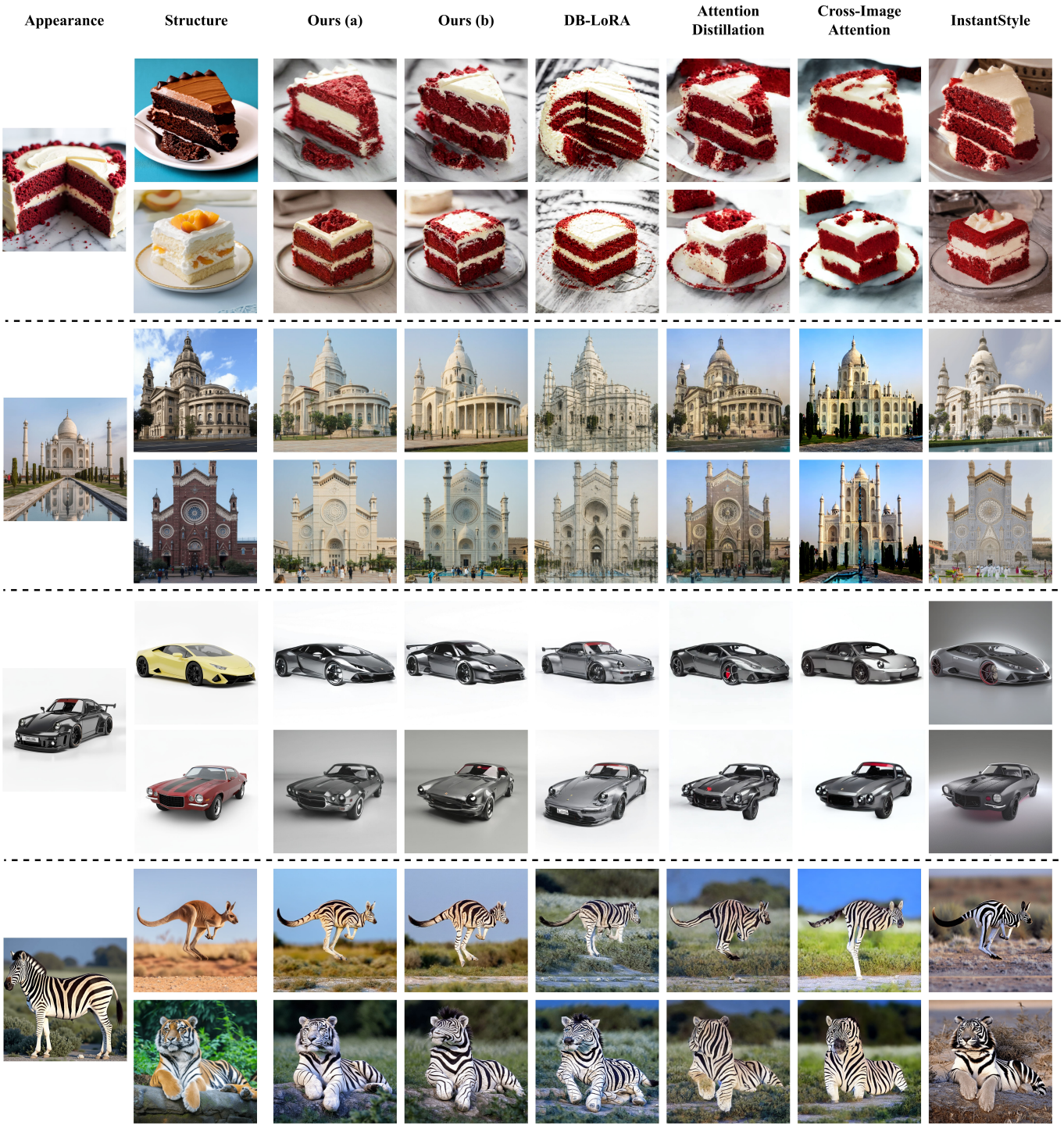

Fig. 18. Qualitative comparison on the appearance transfer task.