Table 2. Shared materials and licenses in this work.

Materials	Licenses
StyleBooth Homepage	-
StyleBooth Dataset	Apache-2.0
StyleBooth Model	Apache-2.0
StyleBooth Model Inference Code	Apache-2.0

A. Overview

In the appendix, we present more implementation details of StyleBooth Dataset in Appendix B including image pairs and textual instruction templates. We also provide the evaluation of our training data in Appendix B.4. Secondly, we explain the experiment implementation details in Appendix C. We show our model's style composition and interpolation ability and more results in Sec. 5 and Appendix D. Furthermore, we discuss social impacts and licence of assets in Appendix E. Finally, we explain the details of human evaluation in Appendix F. We public the related materials as shown in Tab. 2.

B. Dataset Details.

B.1. Style Categories.

The StyleBooth dataset consists of 63 style categories and the image pair numbers of each style are shown in Fig. 8. While some styles have fewer image pairs, the majority is relatively uniformly distributed with the number of image pairs being above 100.

B.2. Image Pair Construction.

We use stylize prompt expansion formats provided by Fooocus [11], see the second column of Tab. 3 for examples. A placeholder of "{prompt}" is set for the original prompt. To generate $BatchA$, we use 217 various prompts to generate diverse style images. Similarly, we select 200 image captions from LAION Art [44] dataset as prompts to generate $BatchB$. 5 samples of prompts used for $BatchA$ and $BatchB$ generation are shown in Tab. 4. See the first column of Fig. 9 for $BatchA$ samples in different styles and the fourth column for $BatchB$. To produce the derived images $BatchA^n$ and $BatchB^n$, we first train a vanilla version of image editing model using the Instruct-Pix2Pix [4] training data based on a pre-trained T2I diffusion model at 512×512 resolution. Each Style Tuner and De-style Tuner is tuned on 1 NVIDIA A800-SXM4-80GB GPU for 10000 steps. We use a learning rate of 0.0001 and a small batch size of 4 and the training resolution is set to 1024×1024. Image pairs are randomly resized to $1\times$ – $1.125\times$ training resolution and then center cropped into $1\times$. For usability filtering, we employ a CLIP-based metric and establish upper and lower thresholds of 0.84 and 0.2. From Fig. 9, you can see the evolution of image quality comparing $BatchA^2$ to $BatchA^1$. The final data is the image pairs of $BatchA^2$ and $BatchA$ which is excellent as a style editing data.

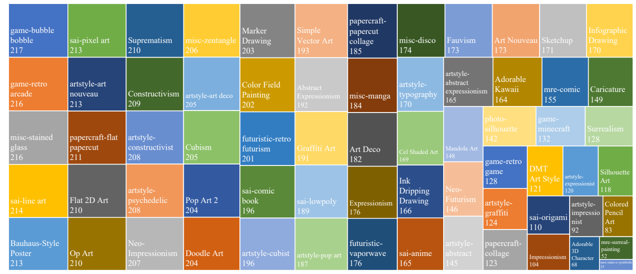

Figure 8. **Style distributions of StyleBooth dataset.** We present the name and image pair numbers for each style. Best viewed when zoomed in.

Figure 9. **Additional dataset samples generated in iterative style-destyle editing.** Pair $BatchA - BatchA^1$ and $BatchB - BatchB^1$ are intermediate image pairs, while final image pairs are $BatchA - BatchA^2$.

B.3. Textual Instruction Templates.

We list 5 samples of machine-generated textual instruction templates for text- and exemplar-based style editing in Tab. 5 respectively. We utilize LLM to generate 15 templates per task. In these instruction templates, "⟨style⟩" and "⟨image⟩" are planted as identifiers for textual style name and exemplar image. During training, we randomly choose from these templates to keep the syntax diversity.

Table 3. Top 5 styles where the number of usable image pairs increase the most after iterative style-destyle editing. Numbers are reported in percentage(%).

Style Name	Prompt Expansion Format	$BatchA^1$	$BatchA^2$	Δ
artstyle-psychedelic	"*psychedelic style {prompt} . vibrant colors, swirling patterns...*"	8.76	95.85	87.10
Suprematism	"*Suprematism, {prompt}, abstract, limited color palette...*"	14.75	96.77	82.03
misc-disco	"*disco-themed {prompt} . vibrant, groovy, retro 70s style...*"	5.53	80.18	74.65
Cubism	"*Cubism Art, {prompt}, flat geometric forms, cubism art*"	20.74	94.47	73.73
Constructivism	"*Constructivism Art, {prompt}, minimalistic, geometric forms...*"	23.50	96.31	72.81
Average	-	38.11	79.91	41.80

Table 4. Samples of prompts for style image and plain image.

Prompt Samples for Style Image
"*A man with a beard*"
"*A lizard with red eyes*"
"*A woman carrying a basket on her back*"
"*Snowy night landscape with houses and trees in the snow*"
"*A close up of a little hamster singing into a microphone*"
Prompt Samples for Plain Image
"*Affordable summer destinations hawaii*"
"*Closeup of a spinach and feta cheese omelet*"
"*Sportsman boxer fighting on black background with smoke boxing*"
"*Healthy green smoothie and ingredients - detox and diet for health*"
"*Masaru Kondo collects Ralph Lauren clothing and accessories.*"

Table 5. Samples of text- and exemplar-based instruction templates. We utilize LLM to generate 15 templates for each task. "$\langle$style$\rangle$" and "$\langle$image$\rangle$" are identifiers for textual style name and exemplar image respectively.

Samples of Text-Based Instruction Templates
"*Let this image be in the style of $\langle$style$\rangle$*"
"*Please edit this image to embody the characteristics of $\langle$style$\rangle$ style.*"
"*Transform this image to reflect the distinct aesthetic of $\langle$style$\rangle$.*"
"*Adjust the visual elements of this image to emulate the $\langle$style$\rangle$ style.*"
"*Reinterpret this image through the artistic lens of $\langle$style$\rangle$.*"
Samples of Exemplar-Based Instruction Templates
"*Please match the aesthetic of this image to that of $\langle$image$\rangle$.*"
"*Adjust the current image to mimic the visual style of $\langle$image$\rangle$.*"
"*Edit this photo so that it reflects the artistic style found in $\langle$image$\rangle$.*"
"*Transform this picture to be stylistically similar to $\langle$image$\rangle$.*"
"*Recreate the ambiance and look of $\langle$image$\rangle$ in this one.*"

B.4. Data Evaluation.

In Tab. 3, we list the top styles where the usability improve the most after one editing-tuning round. The usability metric is also based on CLIP-score, which is the same with that of usability filtering. As we explained before, comparing to the results of vanilla de-style, the average usability rate is increased dramatically from 38.11% to 79.91% by 41.80%. It shows that our Iterative Style-Destyle Editing is effective for quality improvement of paired images. Additionally, we show more visual results and comparison between $BatchA$, $BatchB$ and $BatchA^1$, $BatchA^2$, $BatchB^1$ in Fig. 9, from which we can visually observe the differences.

C. Experiment Implementation Details.

We utilize 8 NVIDIA A800-SXM4-80GB GPUs for our experiments. During inference, we implement classifier-free guidance for both image and text conditions. Following the recommendation in [28], we also apply a re-scaling factor of 0.5 to the outcomes. For instruction-based style editing, we fine-tune the pre-trained model [4] for 5000 steps under 0.0001 learning rate. For exemplar-based style editing, we only tune the alignment layers W and the U-Net [42] decoder under the learning rate of 0.00001 for 35000 steps. Both are trained using Adam [31] optimizer. The scale weighting is only applied during inference in compositional style editing tasks. We adjust scale factors in the range of $[0.5, 1.5]$.

D. Additional Results.

As shown in Fig. 10, we present more qualitative results in Emu Edit benchmark, including comparisons with Emu Edit [45], InstructPix2Pix [4] and Magic Brush [53]. In Fig. 11, we show additional results of exemplar-based style editing with real world images. We use two original images: the David by Michelangelo and the Eiffel Tower, five style exemplars: an animate film stage photo, a Fauvism painting by Henri Matisse, a Cubism painting by Pablo Ruiz Picasso, a post-Impressionist painting by Georges Seurat and a pixel game character. In Fig. 12, we demonstrate an editing result in a compositional style of 3 different styles. Both original image and the style exemplars are real world images. The art work "*The Son of Man*" by Rene Magritte is used as the original image and 3 exemplar images in different style are provided. We show the results under different scale weights.

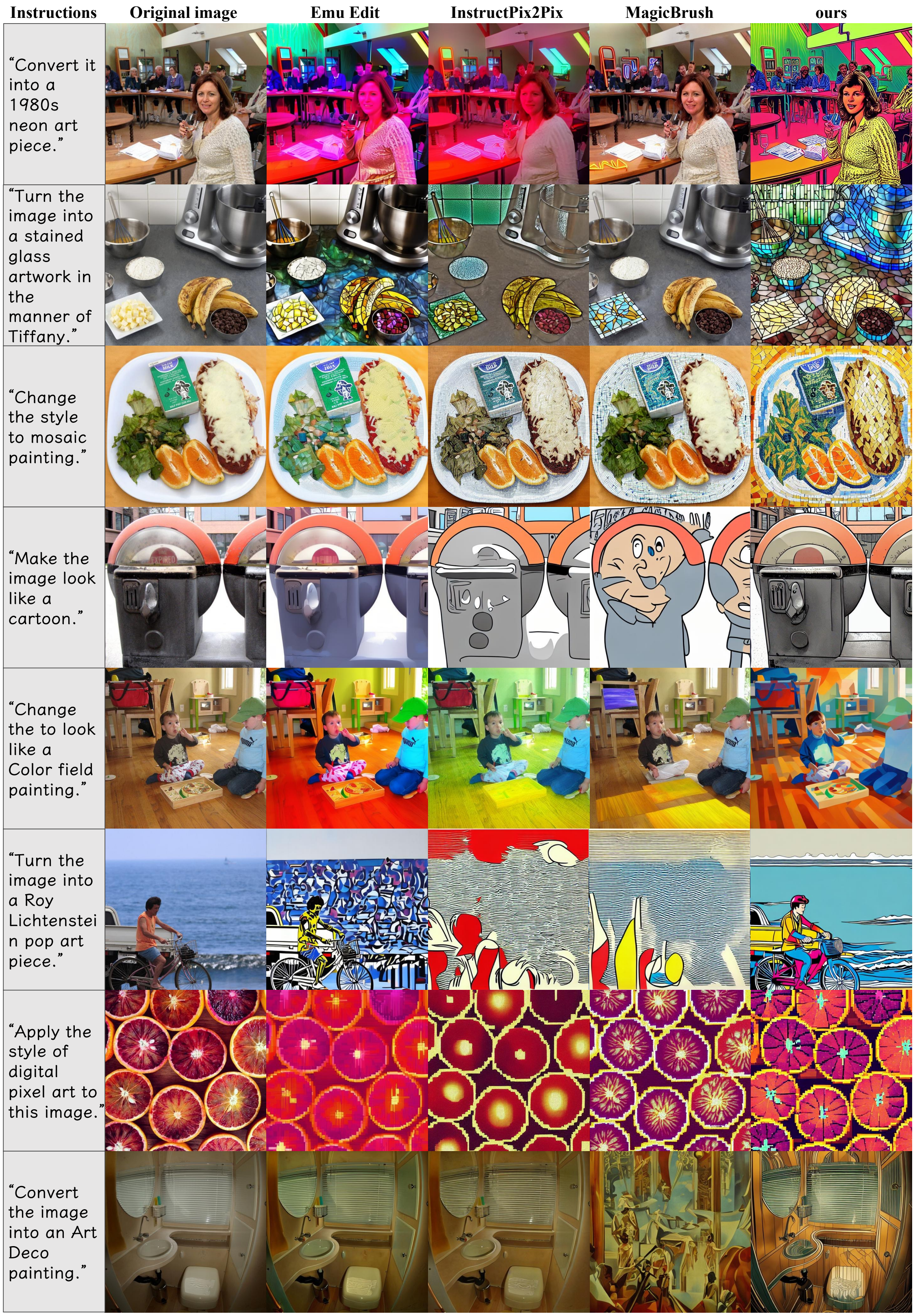

Figure 10. **Additional results comparing with baselines in Emu Edit benchmark.**

Figure 11. **Exemplar-based style editing with real world images.** We present the results of 2 original images in the styles of 5 different art works.

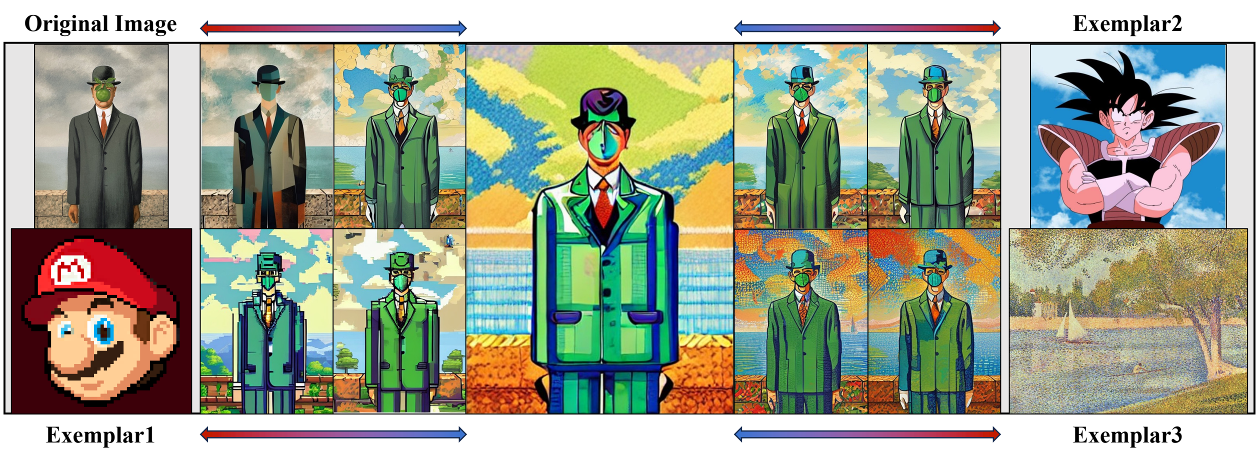

Figure 12. **Compositional style editing combining 3 different styles.** Both original image and the style exemplars are real world images.

E. Discussions.

E.1. Social Impacts.

StyleBooth enables the facile editing of images through simple instructions, providing users with multimodal input options to manipulate images according to their preferences. This approach leads to some positive social impacts, as it allows users to achieve style transfer effortlessly without the need for professional editing tools, substantially enhancing productivity. Moreover, users can explore more liberating style options using text prompts or reference images, thereby offering a creatively enriched editing tool. However, given the inherent risks associated with generative image synthesis, such as malicious use and dissemination, it is imperative to incorporate additional safeguards during the development of such systematic tools.

E.2. License of Assets.

For baselines, Instruction-Pix2Pix [4] inherits this license as it is built upon Stable Diffusion. Magic Brush [53] are released under Creative Commons Attribution 4.0 License. RIVAL [57], StyleAligned [18] and VCT [8] are under Apache-2.0 license.

For datasets, Emu Edit [45] and LAION Art [44] are un-

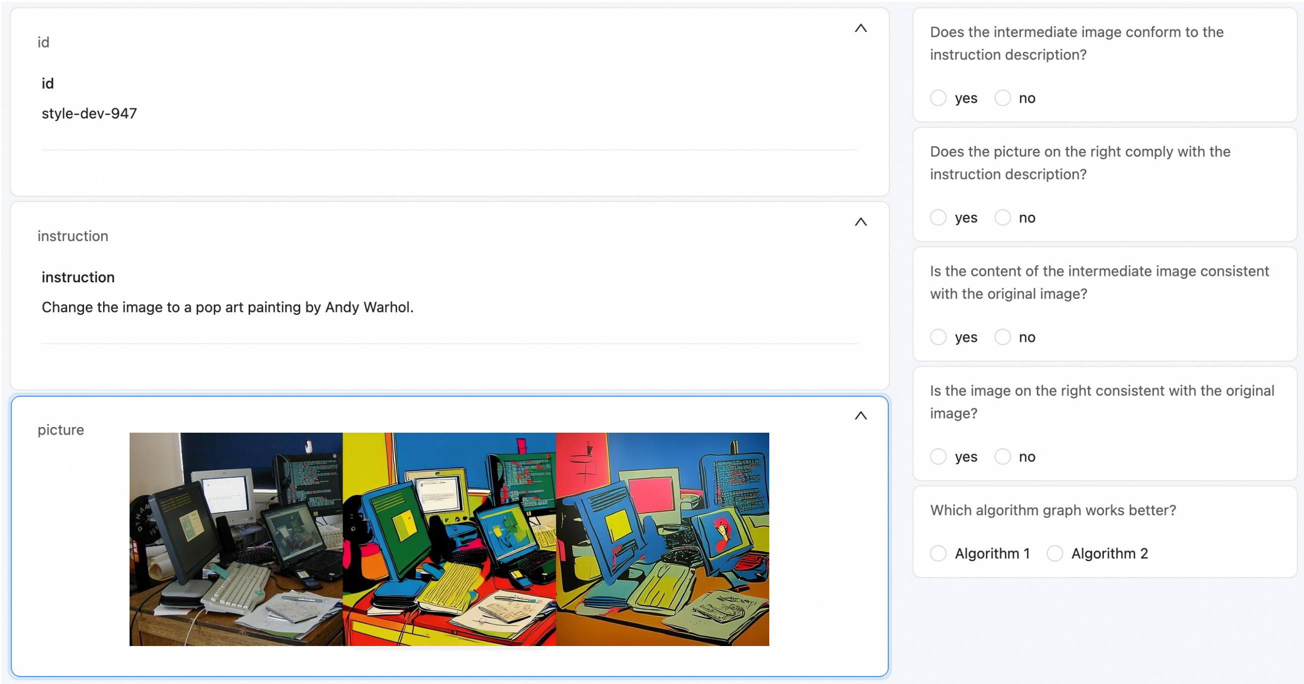

Figure 13. **Screenshots of our anonymous questionnaires for human evaluation.**

der Creative Commons Attribution 4.0 License. According to Stable Diffusion-XL [1], which is under Open RAIL++-M License, we have the right to distribute the generated images for research purpose.

F. Human Evaluation.

We conducted human evaluation to assess the baseline comparisons. We distributed anonymous questionnaires with 5 questions for each editing case, along with instruction, original image, edit result from Method 1 and Method 2 as showed in Fig. 13. The first 2 questions are about whether styles of the 2 editing results are correct and match instructions. The third and fourth questions estimate the content consistency of the 2 methods. At last, the candidates were asked to choice the better one between these 2 methods.