Multimodality-guided Image Style Transfer using Cross-modal GAN Inversion

Hanyu Wang[1]†, Pengxiang Wu[2], Kevin Dela Rosa[2], Chen Wang[2], Abhinav Shrivastava[1]
[1]University of Maryland, College Park [2]Snap Inc.
hywang66@umd.edu {pwu,kevin.delarosa,chen.wang}@snapchat.com abhinav@cs.umd.edu

Supplementary Material

A. Implementation Details

For all of our cross-modal GAN inversion experiments, we utilize a pretrained StyleGAN3-T model [4] that was trained on the WikiArt dataset[1]. We use this[2] implementation in all our experiments. It is worth noting that the performance of this StyleGAN3-T model may be restricted by its pretraining dataset, which only involves the WikiArt dataset. Nevertheless, despite using this domain-limited StyleGAN3-T, our approach still remains competitive, as evidenced by our qualitative findings and user study. By employing more powerful generators, our approach can achieve even better performance.

For cross-modal GAN inversion, we use Adam optimizer [6] with a learning rate of 0.2. We set the summation of all style weights $\{\alpha_i^I\}_{i=1}^{N_I}$ and $\{\alpha_i^T\}_{i=1}^{N_T}$ to be 1000.

To make fair comparisons with previous works, we set the spatial resolution to 512×512 for all image data in our framework. We use a patch size of 256 in all patch-wise CLIP losses. To compute the proposed style-specific CLIP loss, we use the CLIP ViT-B/32 [12] model. Following [2, 5, 7, 11], we apply prompt augmentation [12] to all text descriptions by default. When computing this loss, we resize all inputs to 224×224 to make them compatible with the image encoder of CLIP. Following [7], we apply random perspective augmentation with a distortion scale of 0.5 to all image data used in our main results. In other words, the $\mathrm{aug}(\cdot)$ function defined in our main paper is implemented as RandomPerspective(fill=0, p=1, distortion_scale=0.5) using $\mathrm{torchvision.transforms}$. It takes 20 iterations to run our cross-modal GAN inversion. Our complete code will be made available.

B. Style Text Descriptions and Content Images

We use 11 content images and 20 style images released by [3]. We also use 50 square-shaped images randomly sampled from COCO test set [8] as a supplement to our content set. In addition, we manually collect 44 style text descriptions, including those used by [7]. We list all style text descriptions in the attached file, *style_text.txt*. And we put all content images in the *content* folder.

C. User Study Design

In our user studies, we ask professional annotators from Scale AI[3] to evaluate all our results.

In the main user study (*i.e.*, Table 2 in the main paper), we apply 44 distinctive text-described styles to 61 different content images, giving 2,684 stylized images. For each of them, we ask 10 different annotators to evaluate it from three aspects: style consistency, content preservation, and overall quality. For each aspect, annotators are asked if the stylized image *respects the aspect well* (positive) or *not* (negative). In total we obtain 26,840 responses, where each stylized image received 10 responses.

In our user study for ablation study, we randomly pair the style text descriptions and content images. Specifically, we use all 44 style text descriptions, and pair each of them with 23 randomly sampled content images for style transfer, giving 1,012 stylized images.

In Table 1, we list the number of annotators involved in each evaluation task. As is shown, our user study is based on a sufficiently large number of annotators.

In Figures 5, 6, 7, and 8, we show the annotation user interfaces for evaluating style consistency, content preservation, overall quality, and ablation methods, respectively. In addition, we further show several annotation examples in Figures 9, 10, and 11. The number of positive responses received for these images is listed at the bottom. We observe that the evaluation results from annotators are reasonable and consistent with the quality of stylized images.

D. Style Aggregation Strategy

In Multi-style Boosting (Section 5.1 of the main paper), we propose to aggregate styles $\{\mathcal{S}_i\}_{i=1}^{N_\mathcal{S}}$ to enhance style transfer quality. The aggregation strategy depends on the specific implementation of the IIST method $\mathbf{M}$. Here we

[1]https://www.wikiart.org/
[2]https://github.com/Huage001/AdaAttN
[3]https://scale.com/

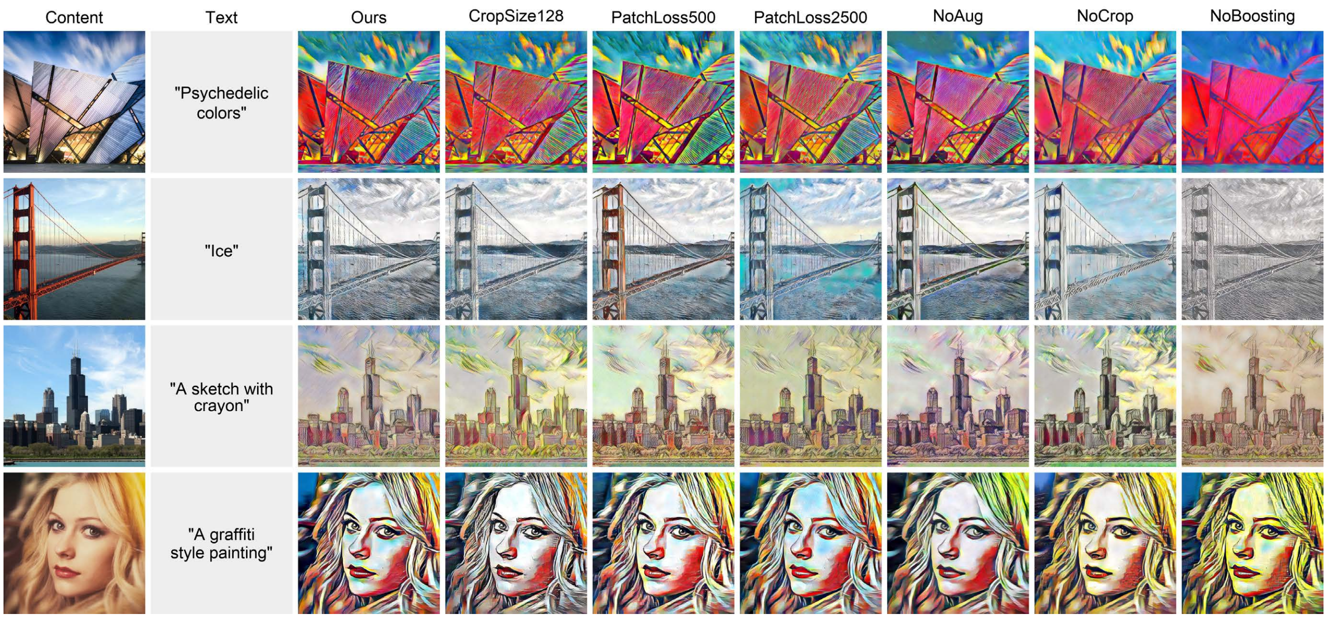

Figure 1. **Qualitative ablation study on different design choices.** We compare our final method with all design choices used in Table 4 of the main paper. Zoom in for a better view.

Table 1. **Number of annotators involved in each evaluation task in our user study.** *Style*, *Content*, and *Overall* are the three aspects in our main user study. *Ablation* refers to all ablation experiments.

Task	Number of Annotators
Style	872
Content	810
Overall	939
Ablation	5041

Table 2. **Additional ablation study on different design choices.** The performance is evaluated through user study. For all these design choices, the user preference percentages are less than 50%, indicating that they are inferior to our method in the main paper.

Setting	Preference % $\uparrow$
Retrieval + AdaAttN	31.1
StableDiffusion + AdaAttN	35.5
ExcludeInv4	44.5
ExcludeInv8	44.9
+ GlobalLoss	49.3

briefly describe the straightforward aggregation algorithm for AdaAttN [9] as an example of $\mathbf{M}$. Similar to many IIST model [3, 10], AdaAttN $\mathbf{M}$ can be decomposed into a feature extraction network $\mathbf{M}_f$ and a style transfer module $\mathbf{M}_t$. Attention mechanism is used for $\mathbf{M}_t$ to process the output features from $\mathbf{M}_f$ in AdaAttN.

After obtaining $\{\mathcal{S}_i\}_{i=1}^{N_\mathcal{S}}$ from cross-modal GAN inversion, we feed them into the feature extraction network $\mathbf{M}_f$ separately and concatenate the outputs together over the sequential dimension at the attention layers in $\mathbf{M}_t$. Since attention layers adaptively focus on the best-matching regions, they can benefit significantly from the high-quality style patterns in the concatenated style representations, while being free from the negative impact of low-quality patterns. The concatenated feature is the F in the Algorithm 2 of the main paper, which is directly used by $\mathbf{M}_t$ to apply style transfer.

E. Latent Initialization

Similar to traditional GAN inversion [1], in cross-modal GAN inversion, the quality of the generated image is sensitive to the initial value of w. Traditional GAN inversion methods often choose the mean latent $\bar{w}$ of the dataset as the initial w. Unfortunately, this initialization is not suitable for our problem as there is no style text description dataset available to compute $\bar{w}$. To overcome this issue, we propose to randomly sample a set of w, from which we pick the best one based on Eq. 3 in the main paper. Formally, we first sample multiple $z_i \sim \mathcal{N}(0, 1)$. Then we run the mapping network of StyleGAN3 on them to obtain $\{w_i\}$. Finally, we calculate

$$\hat{w} = \operatorname*{arg\,min}_{w \in \{w_i\}} L_{\text{sty}}, \tag{1}$$

as the initial value of the StyleGAN3 latent embedding.

Figure 2. **Inverted style representation examples.** The corresponding style text description is displayed above each style representation.

F. Additional Ablation Study

F.1. Qualitative Ablation Study

In order to visually explore the impact of various design choices, we conduct a qualitative ablation study illustrated in Fig. 1. All of the design choices outlined in Table 4 of the main paper are considered. The results indicate that a crop size of 128 (CropSize128) often leads to either over-stylization or under-stylization. Furthermore, the effect of different patch loss weights (PatchLoss500, PatchLoss2500) is negligible, which aligns with the user preference data presented in Table 4 of the main paper. While omitting patch augmentation (NoAug) typically has a minimal effect on the quality of the stylized images, it can sometimes lead to errors such as the incorrect highlighting of edges in the stylized image shown in the first row. In contrast, the omission of patch cropping (NoCrop) can have a more pronounced effect, resulting in oversimplified styles. Finally, our ablation study confirms the importance of multi-style boosting, as performance is significantly degraded when it is not utilized (NoBoosting).

F.2. Additional User Study

We report user study results for additional ablation study. We follow the settings of the ablation user study conducted in our main paper, and consider the text-guided image style transfer task. We report the user preference percentage for the additional design choices in Table 2.

Specifically, we first consider replacing our cross-modal GAN inversion by image retrieval and a text-to-image generative model, respectively, *i.e.*, Retrieval + AdaAttN and StableDiffusion + AdaAttN. For image retrieval, we use CLIP image embedding to retrieve a style representation from WikiArt dataset, which is the same dataset that the StyleGAN3 model was trained on. For the text-to-image generative model, we use the open-source implementation StableDiffusion[4] of the LDMs [13]. We observe that our method significantly outperforms these two design choices, demonstrating the effectiveness of our cross-modal GAN inversion method even if only the text-guided image style transfer task is considered.

Next, we explore if the entire $\mathcal{W}^+$ space is important to ensure the style transfer quality. Inspired by [14, 15], we consider excluding the first 4 layers or 8 layers from the inversion, *i.e.*, ExcludeInv4 and ExcludeInv8. However, we observe that these partial inversion techniques have a negative impact on the style transfer quality.

Finally, inspired by [7], we consider adding a global CLIP loss to the objective function of our cross-modal GAN inversion, *i.e.*, a CLIP loss without image patch cropping. User study result indicates that this additional loss does not improve the user preference percentage. Therefore, we do not add this loss to our main method.

G. Additional Qualitative Results

Inverted Style Representation Examples. Fig. 2 shows some examples of the inverted style representations from style text descriptions. We can observe that many of them do not contain meaningful content, however, they all exhibit certain styles corresponding to the input style text descriptions.

Additional Comparisons with TIST Methods. We show comparison results on more text-image pairs in Figure 12. These examples consistently demonstrate the overall superiority of our method.

Additional MMIST Results. We show more multimodality-guided image style transfer results in Firgure 13. These examples demonstrate how our method combines different styles and faithfully applies them to various content images.

[4]https://github.com/CompVis/stable-diffusion

Figure 3. **Problem of the content loss.** This figure shows randomly picked stylized images and their content loss values. Note that they are obtained from **different randomly selected style transfer methods**. The method names are intentionally hidden to ensure unbiased perception. The fisrt image is the original content image and the remaining ones are the stylized images. Style text descriptions are shown above the images. Content loss values are shown below the images. The highly stylized images (2nd and 4th) appear to incur a higher content loss, even though they largely preserve the original content. In contrast, the 5th stylized image, which deviates minimally from the content image, achieves the best content loss.

Figure 4. **Problem of the content loss.** This figure shows randomly picked stylized images and their content loss values. Note that they are obtained from **different randomly selected style transfer methods**. The method names are intentionally hidden to ensure unbiased perception. The fisrt image is the original content image and the remaining ones are the stylized images. Style text descriptions are shown above the images. Content loss values are shown below the images. The nearly reconstructed stylized image (2nd) consistently achieves the best content loss. More interestingly, the well-stylized image (4th) has a inferior content loss nearly identical to the completely distorted image (1st). This indicates that content loss struggles to differentiate between style variations (4th) and content distortions (1st).

Additional Results of MMIST with Four Style Sources and Cross-modal Style Interpolation. We also show additional MMIST resutls with style interpolation in Figures 14, 15 ,16, and 17. Same as Figure 6 in our main paper, these figures show style interpolation results between 2 text style descriptions and 2 style images. The interpolation ratio for each column or row is fixed to be 1:0, 0.75:0.25, 0.5:0.5, 0.25:0.75, 0:1.

H. Ineffectiveness of Content Loss

Initially, we considered to utilize the content loss employed by CLIPStyler [7] as a metric for quantitative evaluation. However, both our theoretical insights and practical experiments indicated that this content loss doesn not align with human perception.

The content loss as defined in CLIPStyler [7] is calculated as the MSE Loss between the deep VGG features of the stylized image and the content image. Given that VGG is pretrained for recognition tasks, its deep features are acutely sensitive to the distinct visual cues of an input image, such as color and texture. However, variations in color and texture do not necessarily correlate with alterations in the content as perceived by humans. Moreover, modifications in color and texture are the essential outcomes of the style transfer process. This implies that a smaller content loss might indicate a less effective stylization outcome. In the extreme case, an identity mapping function preserves all the content information and has the smallest content loss, but it is a trivial style transfer process and thus undesired. Therefore, while employing content loss during training is

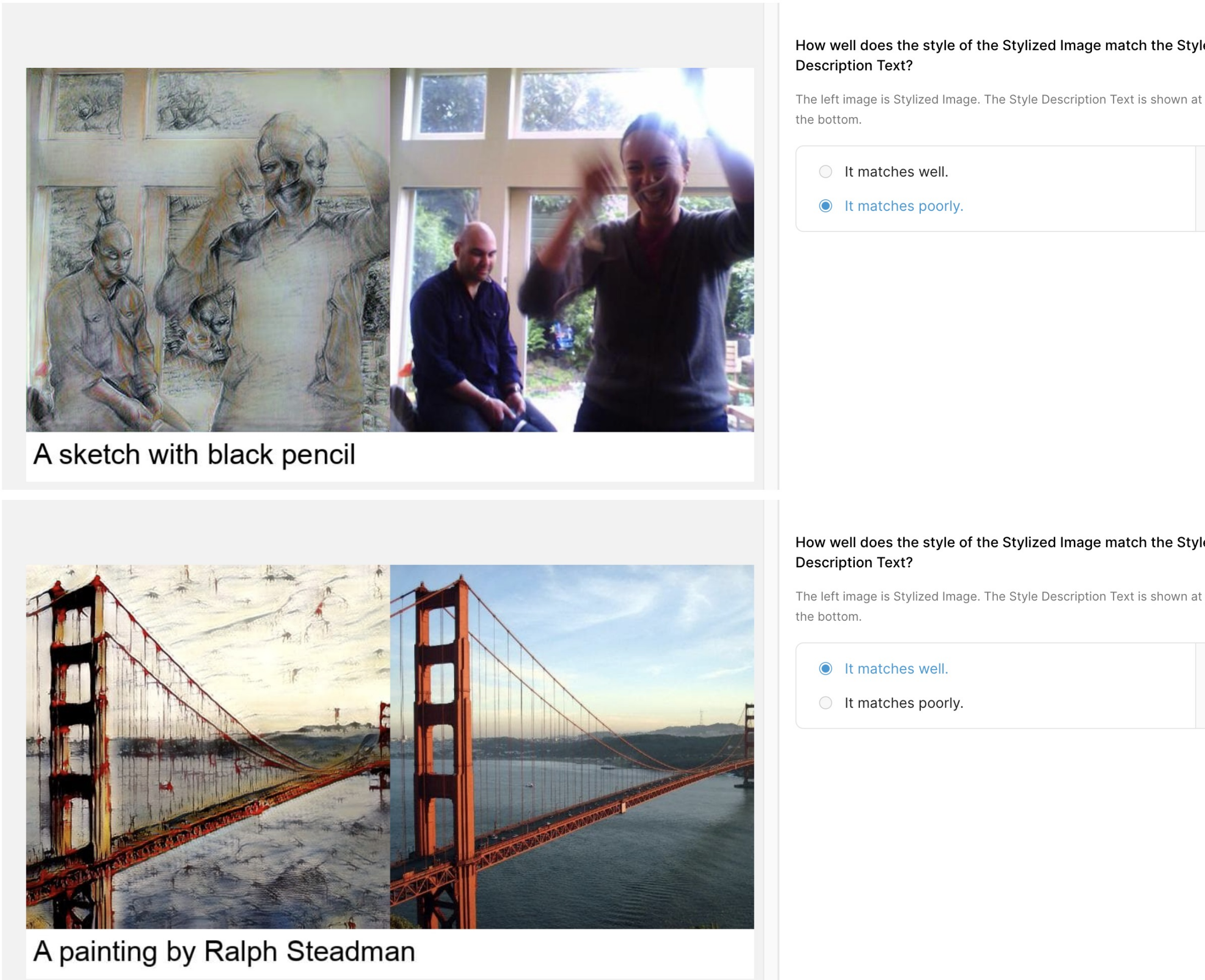

Figure 5. **User interface for image evaluation in user study.** Here we evaluate *Style Consistency*.

not problematic due to the concurrent use of style loss, which ensures the style quality, its application as an evaluation metric is unsuitable.

To further validate our analysis regarding the limitations of the content loss defined in CLIPStyler [7], we randomly pick stylized images from different style transfer methods and compute their content loss for comparison. The results are shown in Figures 3 and 4. Note that the names of selected methods are intentionally hidden to ensure unbiased perception. In Figure 3, the highly stylized images (2nd and 4th) appear to incur a higher content loss, even though they largely preserve the original content. In contrast, the 5th stylized image, which deviates minimally from the content image, achieves the best content loss. In Figure 4, the stylized image (2nd) which nearly reconstructs the original image consistently achieves the best content loss. More interestingly, the well-stylized image (4th) has a inferior content loss nearly identical to the completely distorted image (1st). This indicates that content loss struggles to differentiate between style variations (4th) and content distortions (1st). In summary, the content loss defined in CLIPStyler appears ill-equipped to differentiate between style modifications and content distortions. Therefore, we opt not to use content loss as an evaluation metric in this paper.

I. Limitation

While our approach proves robust across various applications, it is intrinsically constrained by its reliance on the pretrained style representation generator, the adapted IIST method, and the CLIP model, which is utilized to construct the loss function in the cross-modal GAN inversion algorithm.

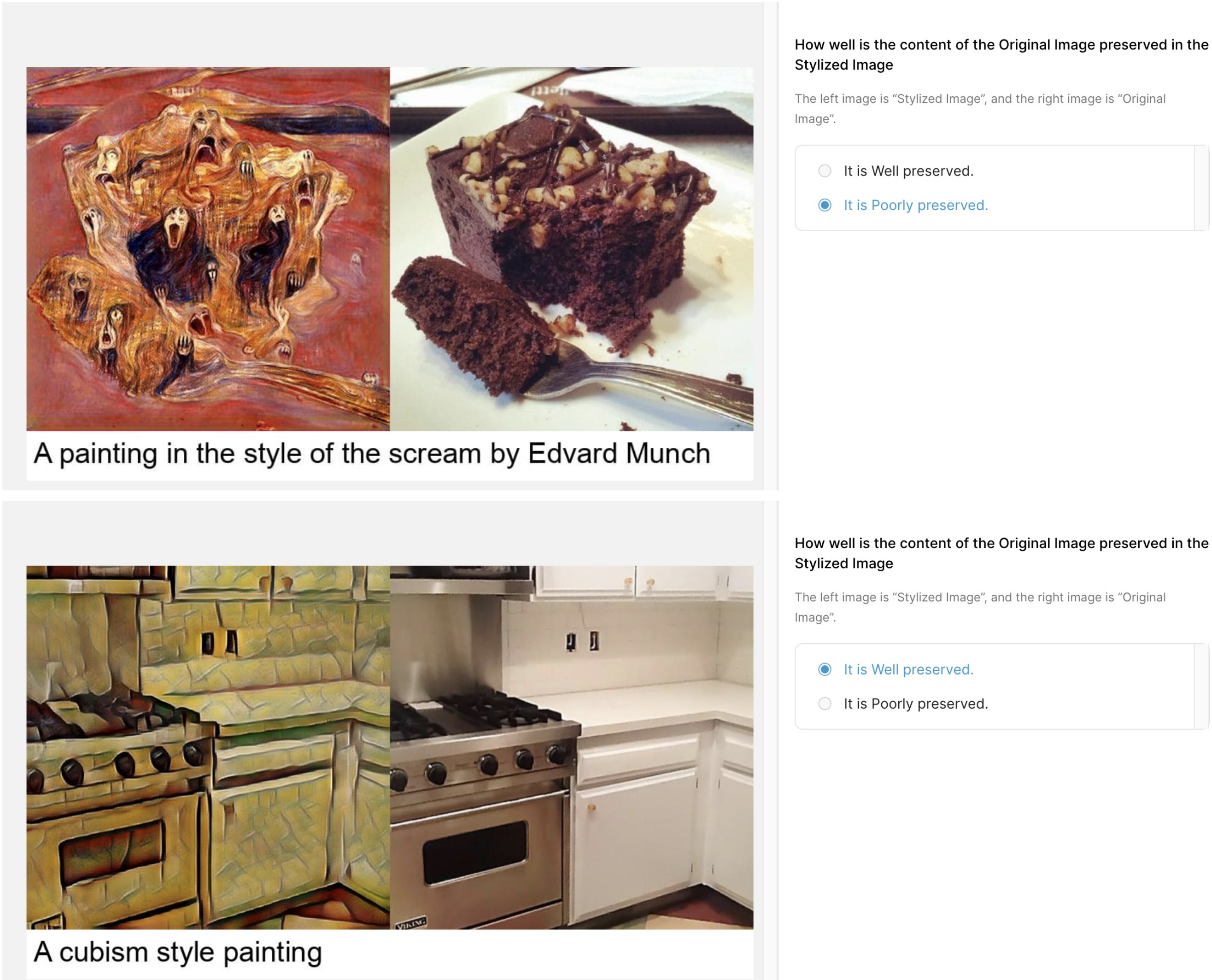

Figure 6. **User interface for image evaluation in user study.** Here we evaluate *Content Preservation*.

We utilize the WikiArt pretrained StyleGAN3 as our style representation generator. While this model encompasses a broad spectrum of styles, its effectiveness can be compromised when confronted with out-of-distribution styles. This limitation arises from the finite scope of the WikiArt dataset. Consequently, when presented with certain styles that are outside this dataset's domain, our style transfer outcomes might not achieve the desired quality.

Similarly, the efficacy of our solution is significantly influenced by the adapted IIST method. This component executes the style transfer after the generation of intermediate style representations. If the adapted IIST method manifests any limitations or biases, it can have a direct negative impact on the results generated by our method.

Moreover, the CLIP model and the Style-specific CLIP Loss are not perfect. Potential inaccuracies in these parts may yield imprecise intermediate style representations, further influencing the quality of the stylized images.

In addition, our method requires a per-sytle optimization procedure for fast style transfer. However, this optimization can be time-intensive, potentially hindering our method's application in time-sensitive scenarios. An alternative could be training a feed-forward style transfer network to eliminate the need for per-style optimization. We leave this potential improvement direction as future work.

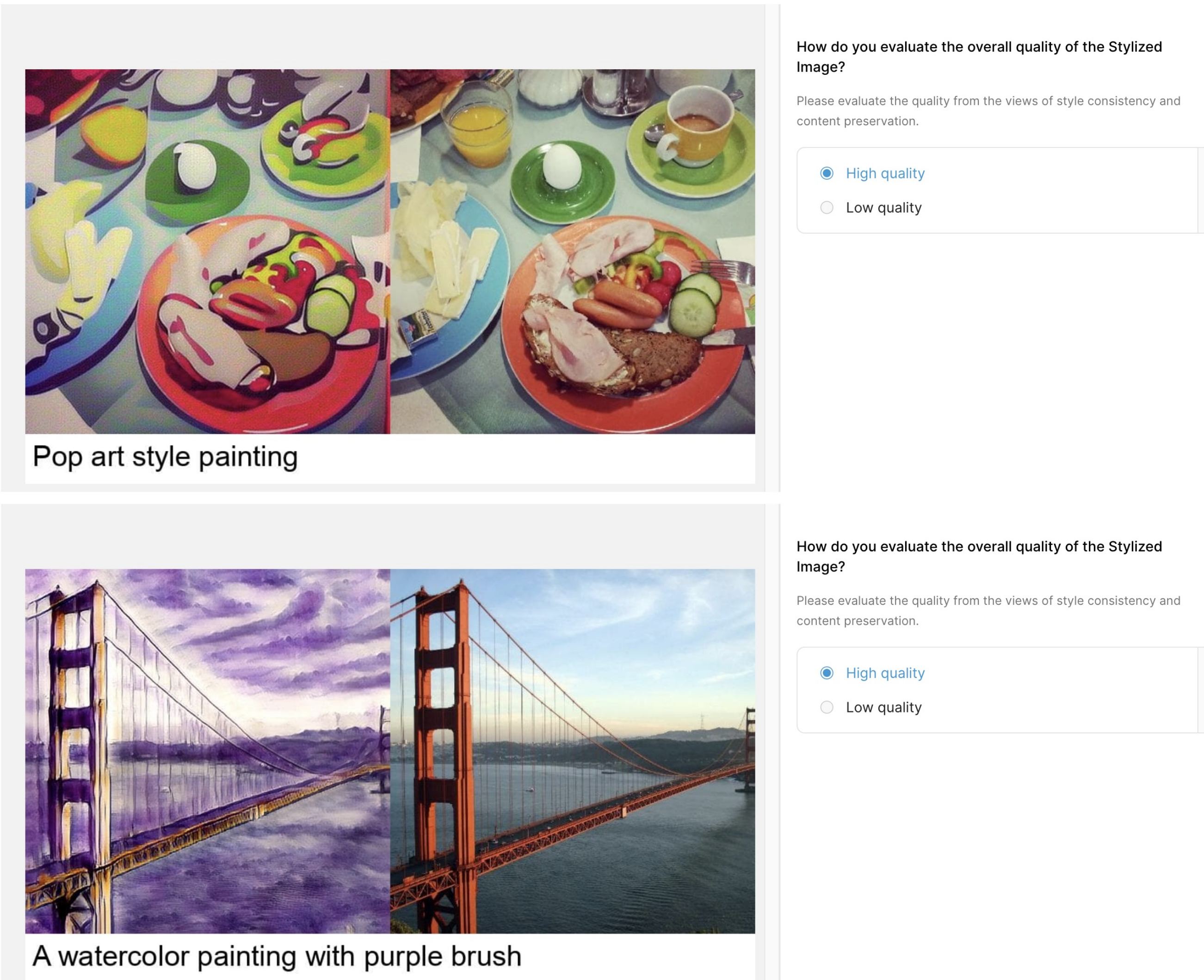

Figure 7. **User interface for image evaluation in user study.** Here we evaluate *Overall Quality*.

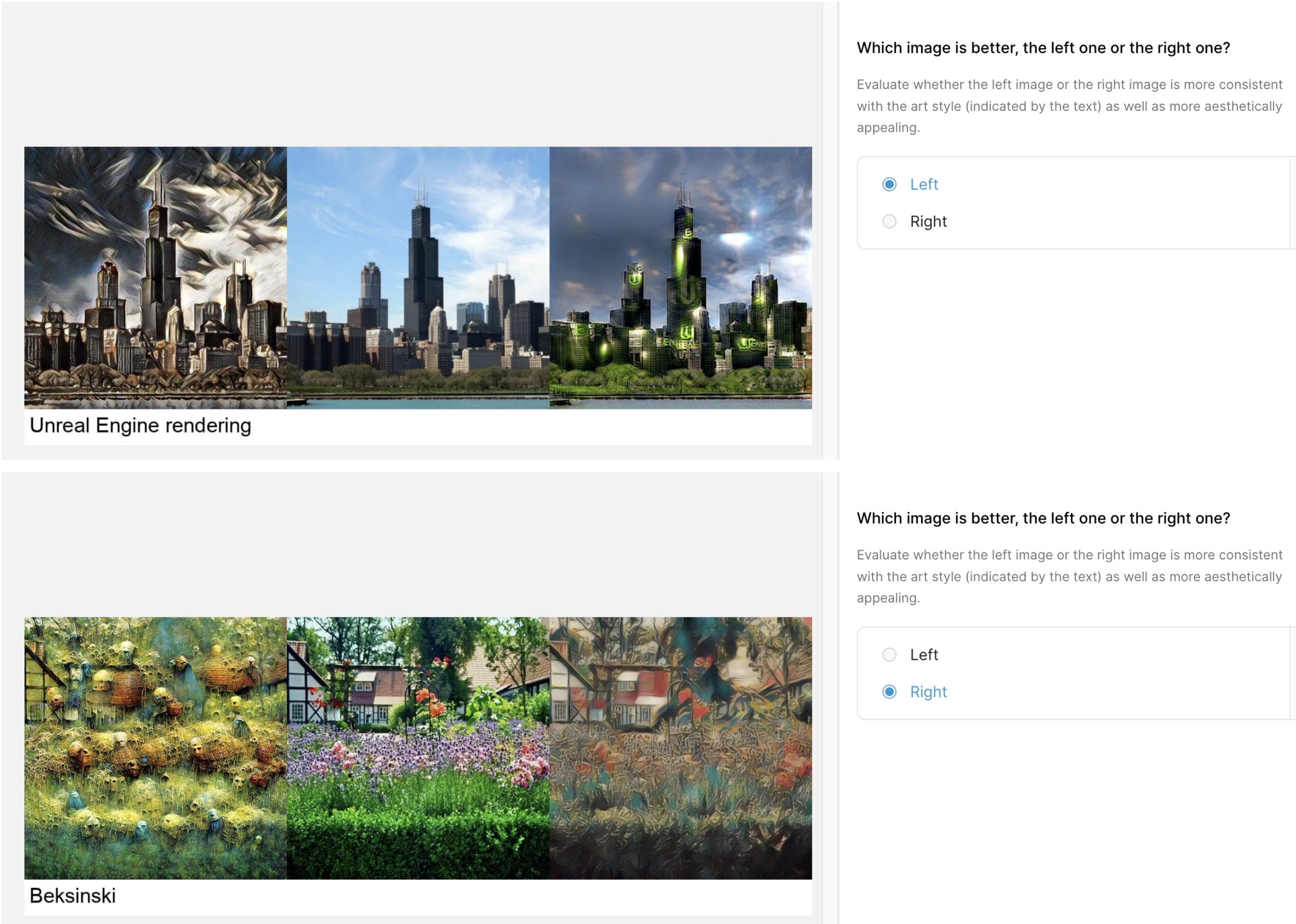

Figure 8. **User interface for image evaluation in user study.** This user interface is used for ablation studies.

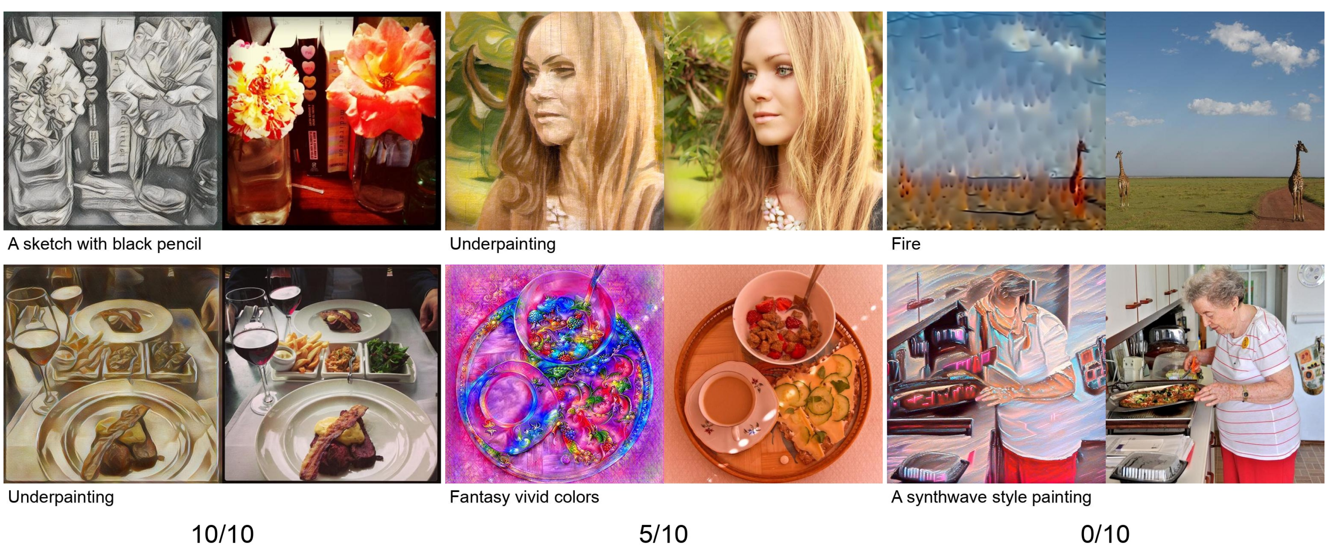

Figure 9. **Examples of *Style Consistency* annotations.** At the bottom of each column we show the number of positive responses received over the total response number.

Figure 10. **Examples of *Content Preservation* annotations.** At the bottom of each column we show the number of positive responses received over the total response number.

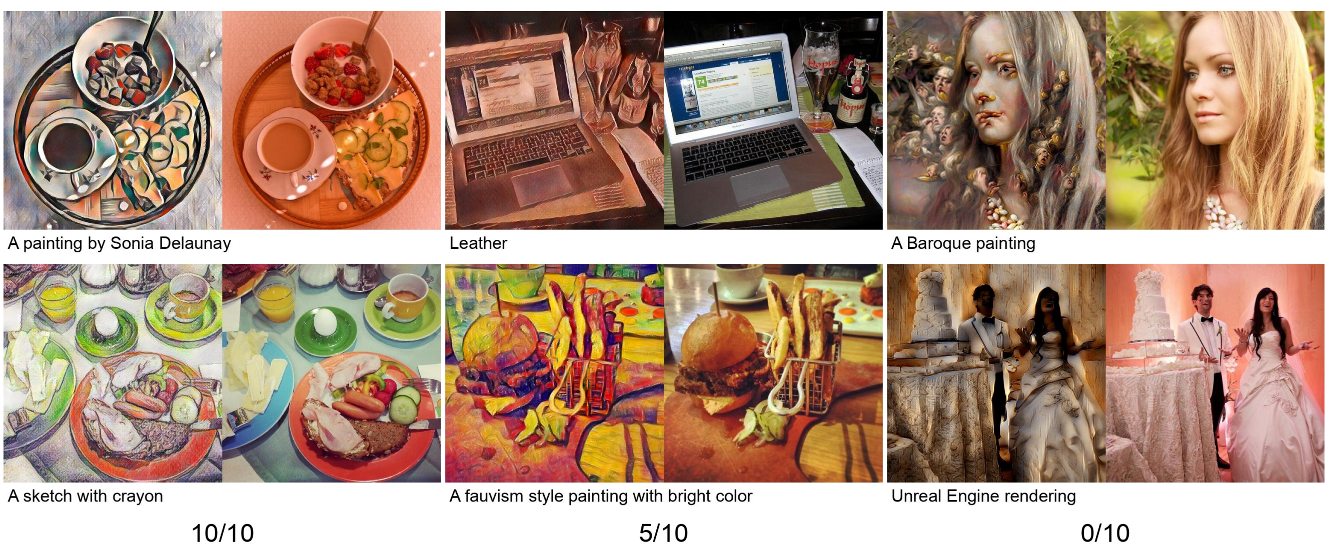

Figure 11. **Examples of *Overall Quality* annotations.** At the bottom of each column we show the number of positive responses received over the total response number.

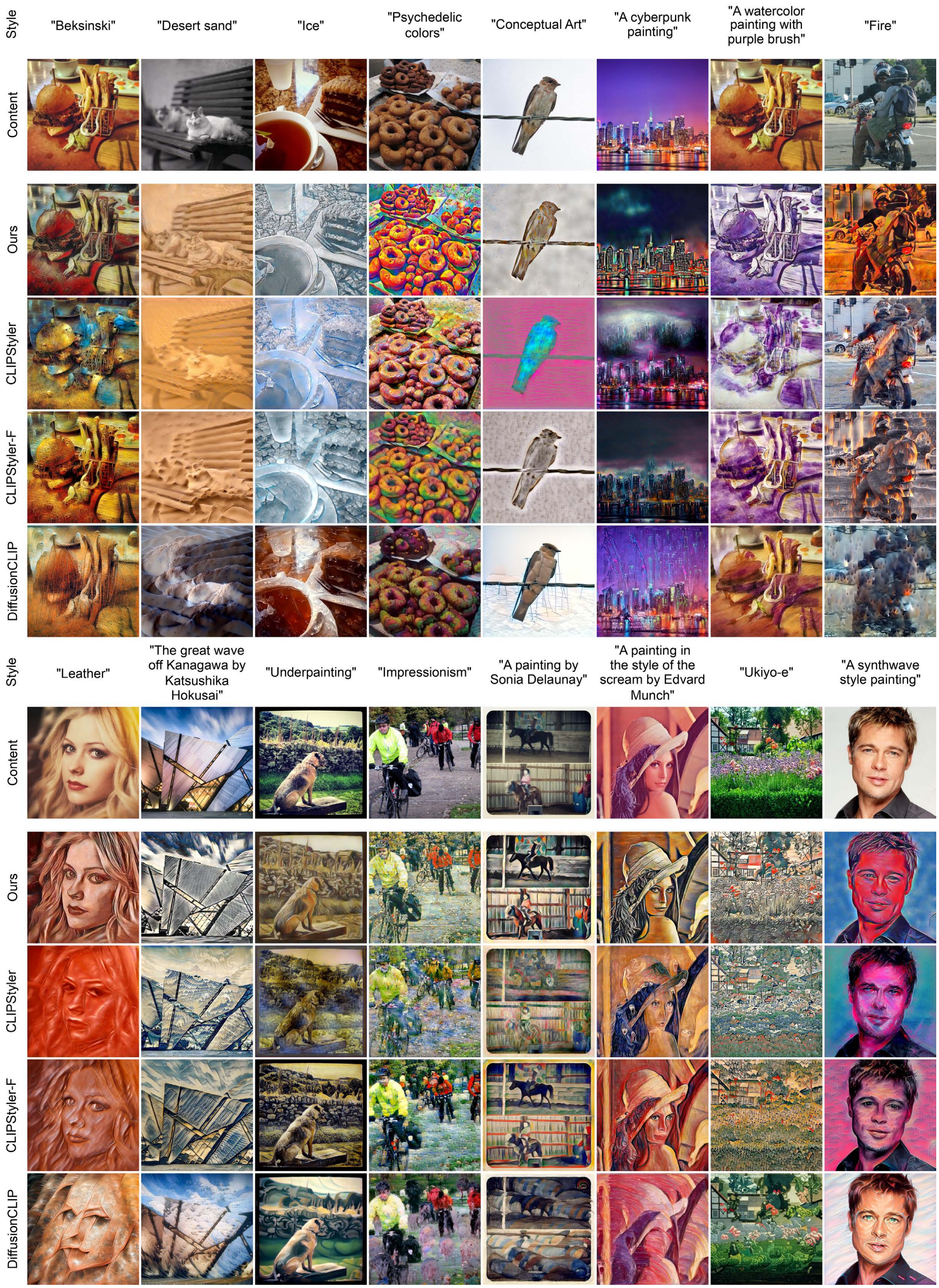

Figure 12. **Additional comparison with other TIST methods.**

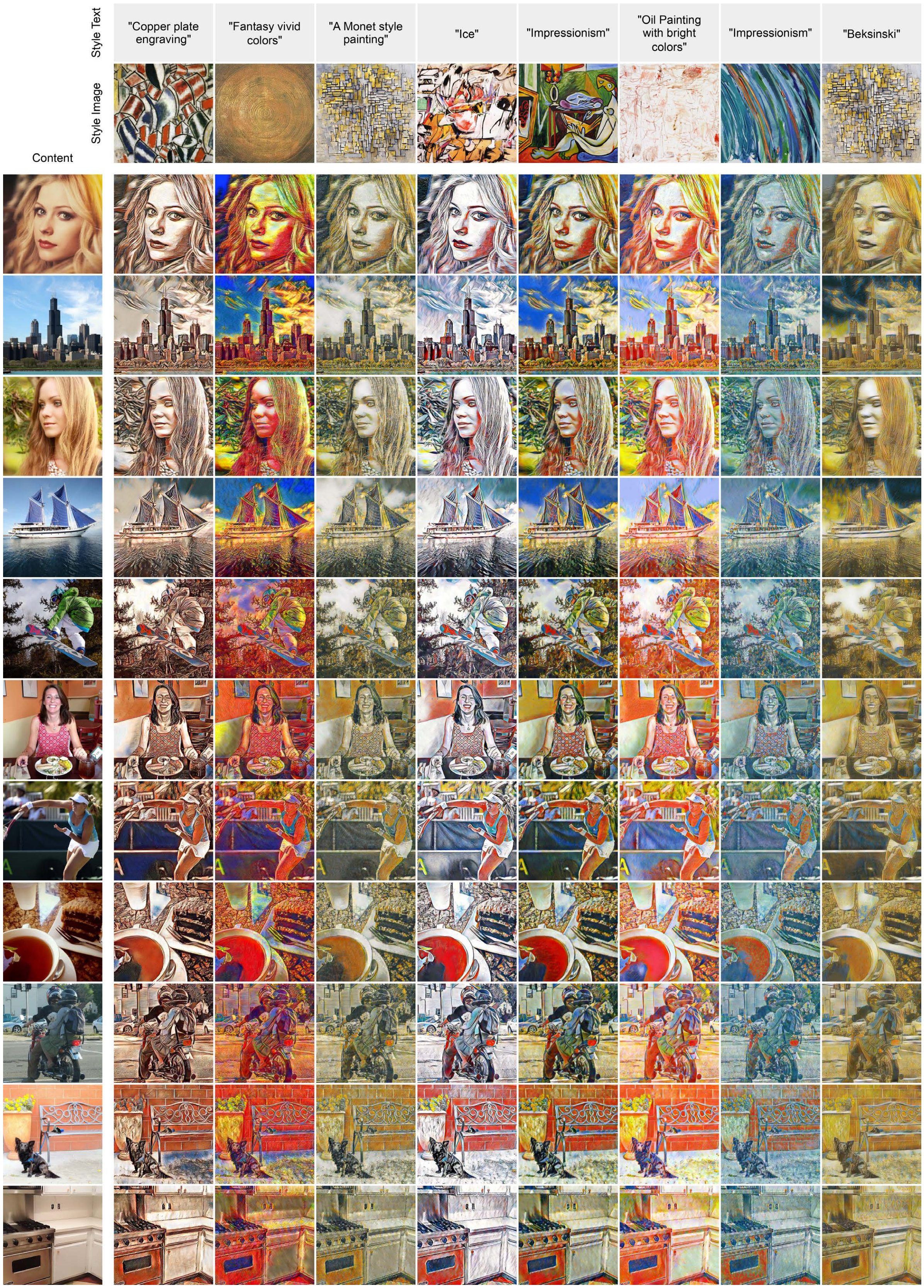

Figure 13. **Additional MMIST results.**

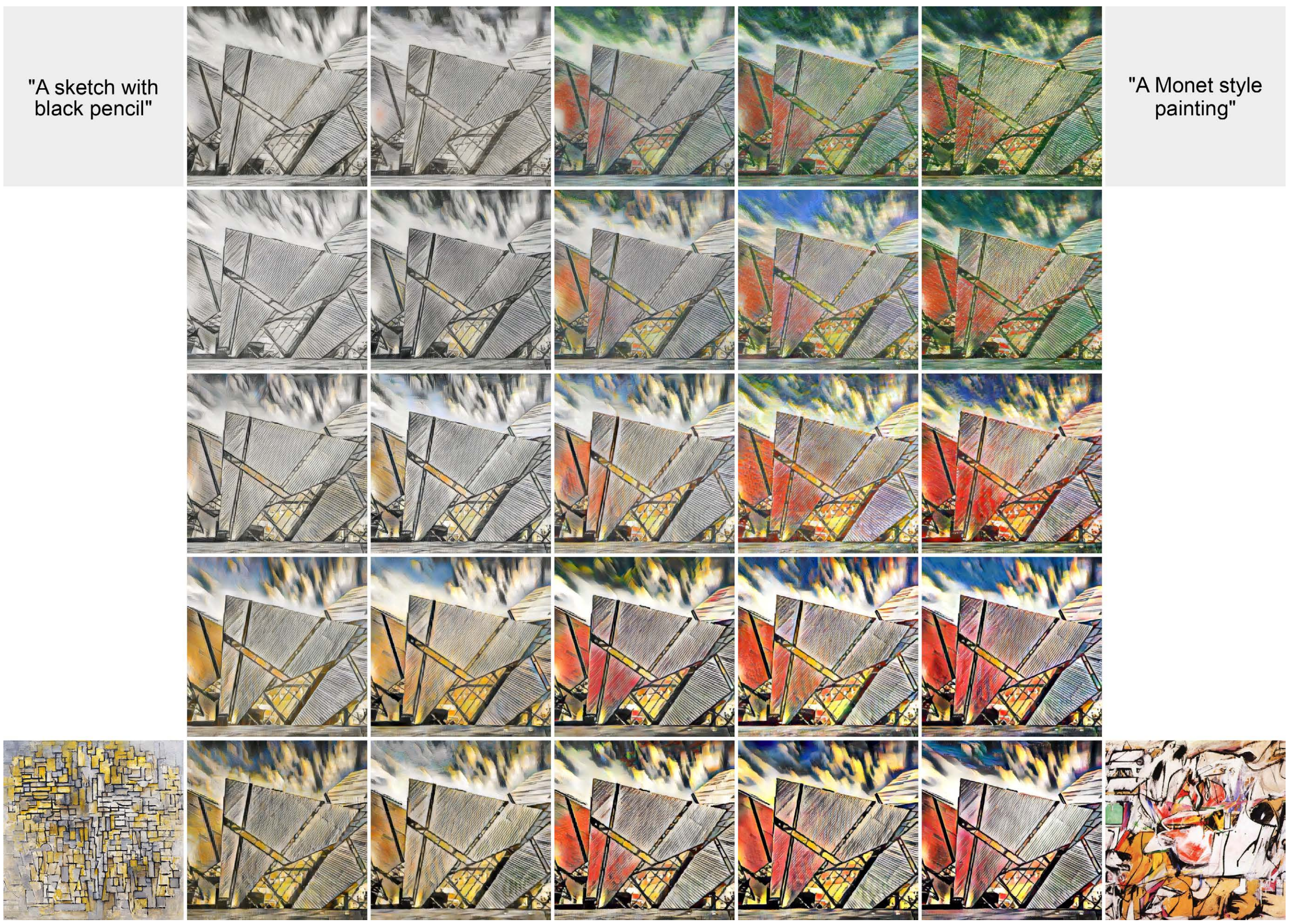

Figure 14. **Additional MMIST results with four image and text styles and style interpolation.** (1)

Figure 15. **Additional MMIST results with four image and text styles and style interpolation.** (2)

Figure 16. **Additional MMIST results with four image and text styles and style interpolation.** (3)

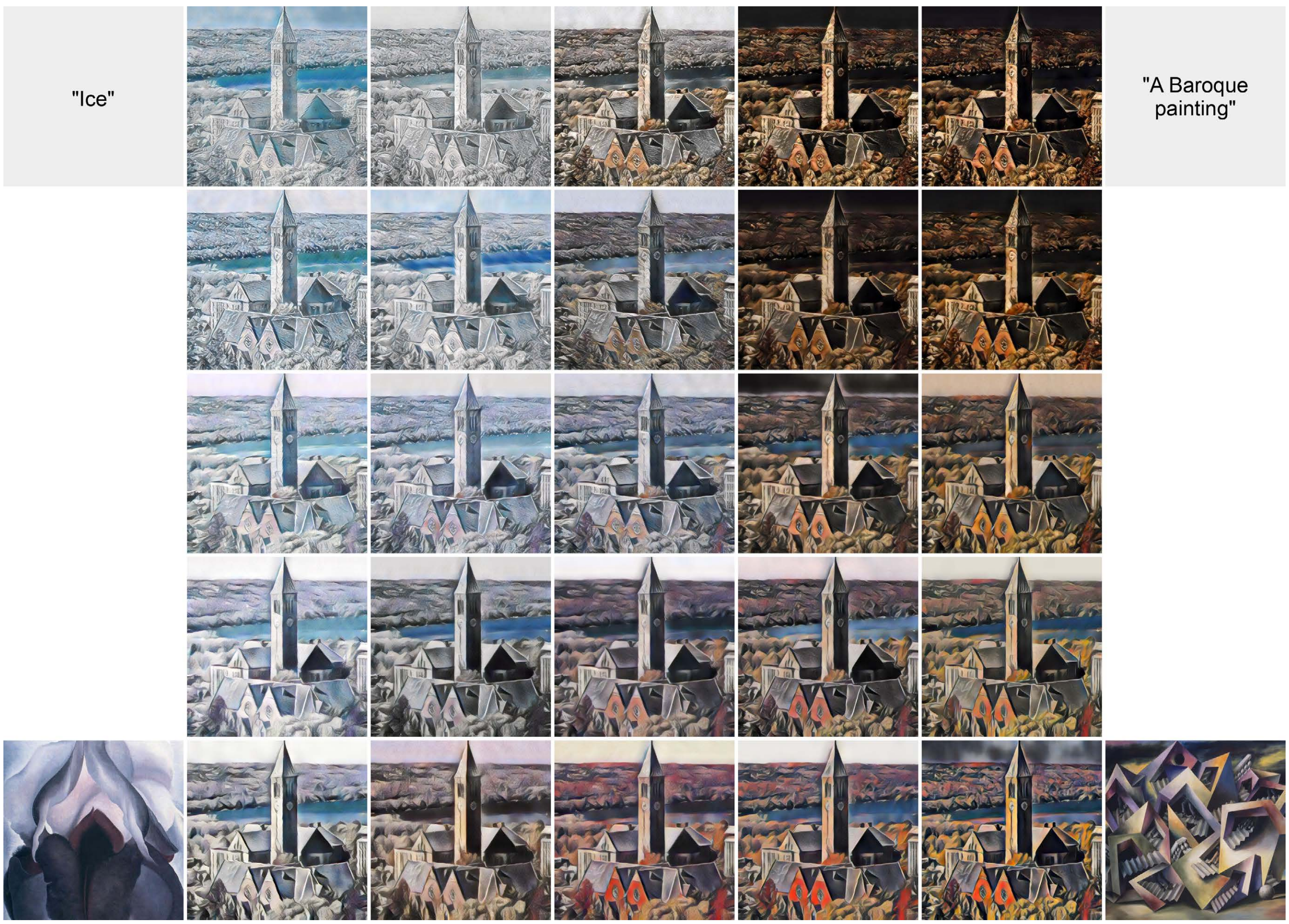

Figure 17. **Additional MMIST results with four image and text styles and style interpolation.** (4)

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 2

[2] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021. 1

[3] 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*, pages 1501–1510, 2017. 1, 2

[4] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 1

[5] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 1

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[7] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 1, 3, 4, 5

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[9] 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*, pages 6649–6658, 2021. 2

[10] 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*, pages 5880–5888, 2019. 2

[11] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 1

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[14] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021. 3

[15] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Towards open-world text-guided face image generation and manipulation. *arXiv preprint arXiv:2104.08910*, 2021. 3