

## A. Appendix

### A.1. Implementation Details

The StylerDALLE NAT model consists of a 4-layer encoder and an 8-layer decoder while the attention head number is 8 and the hidden dimension is 512. We use PyTorch [26] to implement our method. To train our model, we use the train-set of COCO [21] dataset, which contains 82,783 images while each image has 5 captions. In the self-supervised pre-training stage, we only use the images in the COCO train-set. We train the NAT model for 25 epochs with a learning rate of  $1e-4$ . We use Adam [16] optimizer. In the style-specific fine-tuning stage, we use both the images and the captions. In particular, we utilize all the caption annotations to enhance the model robustness, as usually human annotates different captions of a single image. Notably, the caption annotations are only used at the fine-tuning stage. In other words, StylerDALLE does not need to use image caption as input at inference time. We only fine-tune the decoder of the NAT model, keep the encoder frozen. We use Adam optimizer with a learning rate of  $1e-6$ . For CLIP model, we use the CLIP ViT-B/32 model<sup>1</sup>. For pretrained vector-quantized image tokenizer, we use the officially released dVAE of DALL-E<sup>2</sup> or the VQGAN of Ru-DALLE<sup>3</sup>. For both training stages, the model is trained on single RTX-A6000 GPU for 24 hours.

To compare with CLIPStyler, we use the official implementation.<sup>4</sup> For all reference image-based comparing methods, we use the officially released trained models.<sup>5</sup>

### A.2. Ablation Study

We study two ablations of StylerDALLE: (1) without captions, and (2) without scaling.

Firstly, we ablate the usage of captions in formulating the prompt-based reward during the style-specific fine-tuning stage (Sec. 4.2). In more detail, instead of using the CLIP similarity between the stylized image  $I^s$  and the prompt  $t_q$  (which combines the style description  $t_s$  and the image caption  $t_a$ ) as the reward, we discard  $t_a$  and we compute the CLIP similarity between the stylized image  $I^s$  and the style description  $t_s$  as the reward. As shown in Fig. 7, the two models StylerDALLE-1 and StylerDALLE-Ru show different results on the ablation “w/o captions”. For StylerDALLE-1 (Fig. 7(a)), we see that the results of the full model are slightly better. In the results of StylerDALLE-1, the details are preserved bet-

ter, and the colors are closer to the light and muted colors used in watercolor painting. Moreover, the results are overall harmonious as there are few abrupt brushstrokes. Meanwhile, “StylerDALLE-1 w/o captions” also presents a satisfying style transfer quality, as the results keep a good balance between the stylization and content maintainance. This indicates our method can also work for the dVAE of DALL-E when no caption is provided, thus being less annotation-dependent. Nevertheless, “StylerDALLE-Ru w/o captions” (Fig. 7(b)) fails to keep content consistency, emphasizing the significance of using captions as part of the language supervision in the Reinforcement Learning process for maintaining the content.

Secondly, we ablate the operation of down-sampling as introduced in Sec. 4. Specifically, we directly input the discrete tokens of the full-resolution image to the NAT model while conducting the same self-supervised pre-training and style-specific fine-tuning. As shown in the results of “StylerDALLE-1 w/o scaling” (Fig. 7(a)) and “StylerDALLE-Ru w/o scaling” (Fig. 7(b)), scaling is an important procedure in StylerDALLE: when the NAT model is input with the discrete tokens of the full-resolution image, the style cannot be incorporated effectively through the Reinforcement Learning fine-tuning stage.

### A.3. Inference Time

To generate a single  $256 \times 256$  stylized image (including the time to down-sample, encode, translate through the NAT, and decode), StylerDALLE needs 0.076s, which is the average time computed over the COCO val-set using an RTX-A6000 GPU. The main idea of our paper is to use large-scale pretrained image generative models for style transfer and we focus on using vector-quantization-based methods. Therefore, we conclude that as compared to style transfer methods that are based on large-scale diffusion models, StylerDALLE has the advantage of having less inference time. For instance, as reported in a recent paper [1], Imagen [32] takes 9.1s to generate a  $256 \times 256$  image on TPUv4 accelerators.

### A.4. User Study Details

Other than the quantitative analysis and qualitative analysis, as in Tab. 2, we further involve human subjects to evaluate the style transfer results of StylerDALLE and the comparing method CLIPStyler. To help the participants know the styles, at the beginning of evaluating each style, we incorporate several illustrations of the style (Tab. 8(a)). We show part of the questionnaire in Tab. 8(b). We use Google Forms to collect user opinions.

### A.5. Additional Experimental Results

**Additional Comparison Results.** In Fig. 9, we illustrate the additional comparing results between StylerDALLE

<sup>1</sup>CLIP: <https://github.com/openai/CLIP>

<sup>2</sup>DALL-E: <https://github.com/openai/dall-e>

<sup>3</sup>Ru-DALLE: <https://github.com/ai-forever/ru-dalle>

<sup>4</sup>CLIPStyler: <https://github.com/cyclomon/CLIPStyler>

<sup>5</sup>AesUST: <https://github.com/EndyWon/AesUST>, StyTr2: <https://github.com/diyiyiii/StyTr-2>.



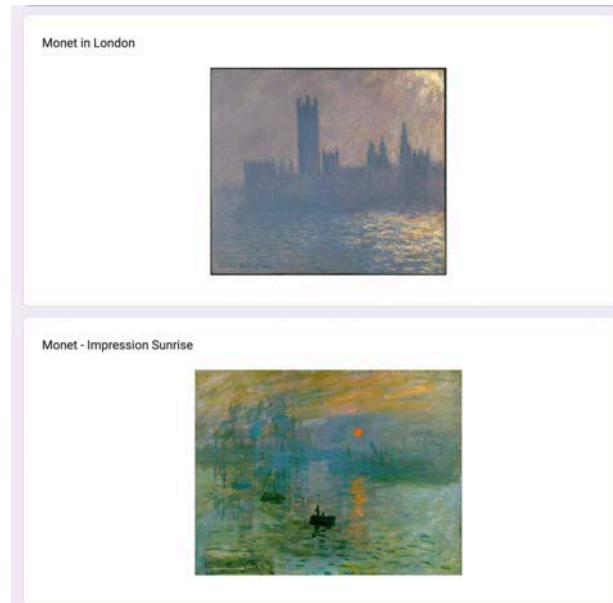
Figure 7: Ablation study on StylerDALLE.

and CLIPStyler-Optimization (i.e., the mainly proposed method in the paper). As shown, CLIPStyler-Optimization suffers from two issues. Firstly, there are many inharmonious artifacts that appear in the stylized images. For example, there are many plant-like artifacts in the stylized results of “Monet” and multiple suns in the “Monet Sun Impression” results. Secondly, the texts related to the language instructions are written in stylized images unexpectedly. For instance, as in the top example of the “fauvism” train, the written text “fauvism” is on the front of the bus.

On the contrary, both StylerDALLE-1 and StylerDALLE-Ru do not have the above two issues. Furthermore, our results achieve well-characterized stylization results consistent with language instructions, and different styles are expressed with varied and distinctive brushstrokes related to the specific style. In the following,




we give more generated results of StylerDALLE-1 and StylerDALLE-Ru.

**Additional Qualitative Results.** We give more stylized results produced by StylerDALLE-Ru in Fig. 10, Fig. 11 and Fig. 12, and StylerDALLE-1 in Fig. 13, Fig. 14 and Fig. 15, respectively. In particular, we also show the intermediate results  $\hat{I}$  (as in Fig. 2), which are generated with the output tokens using the model right after the self-supervised pre-training (and before the style-specific fine-tuning stage). Similar to what we have concluded, both StylerDALLE-1 and StylerDALLE-Ru achieve distinctive and harmonious stylized results on various styles and images. In addition, the differences between  $\hat{I}$  and  $I^s$  are significant. As shown,  $\hat{I}$  is photo-realistic while  $I^s$  presents varied brushstrokes, edges, and colors with respect to each style instruction, indicating that StylerDALLE has been



(a) We illustrate each style with several examples to let the participant know the styles.

1. Which image looks like Monet's artwork most?


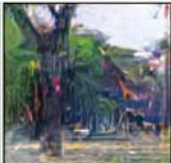





(a)
(b)
(c)

☐ a
 ☐ b
 ☐ c

2. Which image looks like Monet's artwork most?

(a)
(b)
(c)

☐ a
 ☐ b
 ☐ c

(b) In each question, we ask the participant to select one image that is most likely to be of the target style. The order of the candidates is randomly shuffled.

Figure 8: Illustrations of the user study details.

effectively fine-tuned with our language-guided rewards in the Reinforcement Learning stage.

By comparing the results of `StylerDALLE-1` and `StylerDALLE-Ru`, although we draw the joint conclusions as above, we also see the differences between the two, resulting from the usage of different vector-quantized image tokenizers. For example, `StylerDALLE-Ru` achieves clearer stylized images, as it is implemented based on the VQGAN image tokenizer. On the other hand, our method, i.e., `StylerDALLE` has been proven effective on both vector-quantized image tokenizers. It is reasonable to expect that the style transfer results can be further improved by using more advanced vector-quantized image tokenizers if they could be open-sourced.



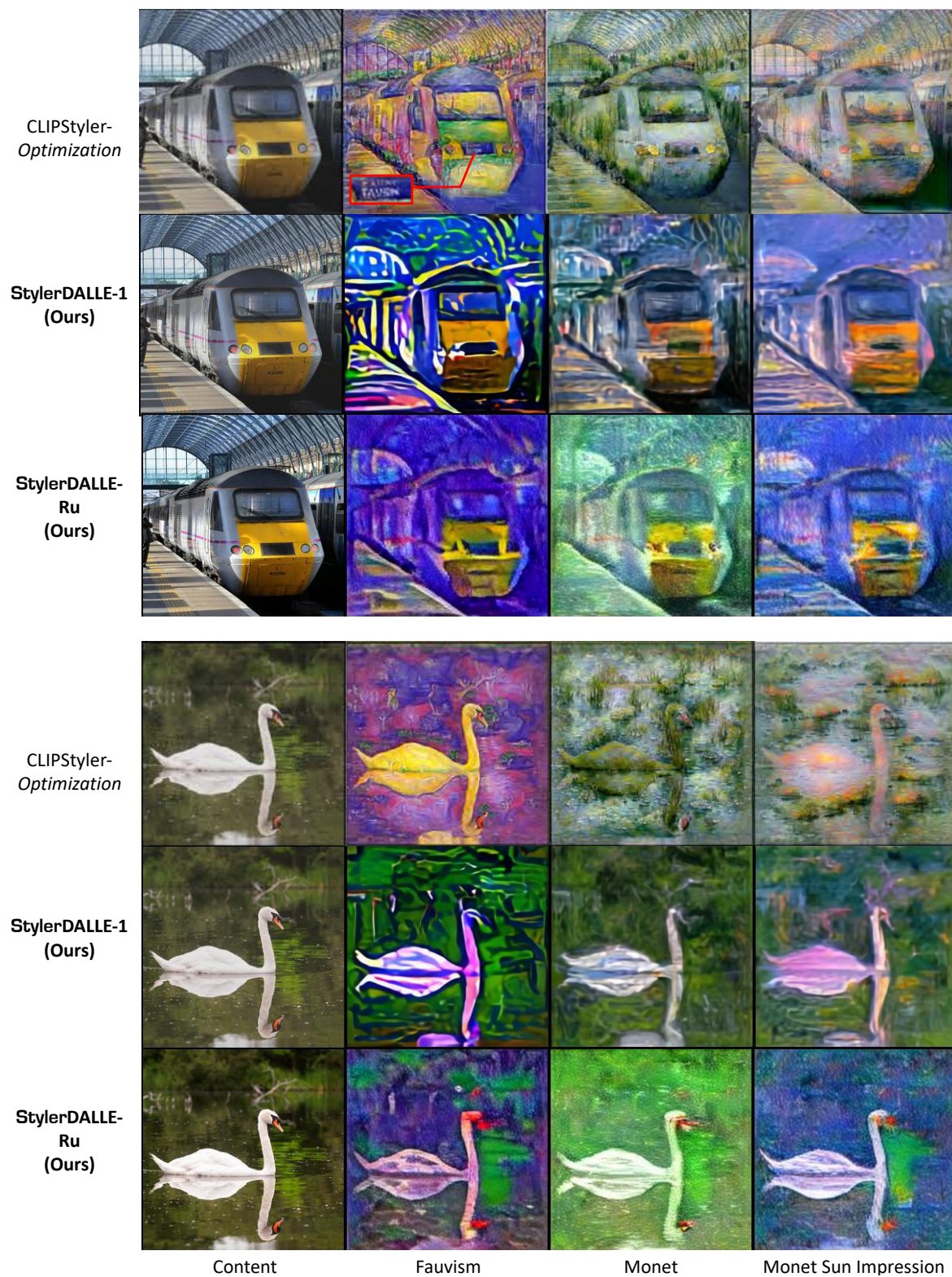


Figure 9: Comparisons between StylerDALLE and CLIPStyler, styles are shown on the bottom.



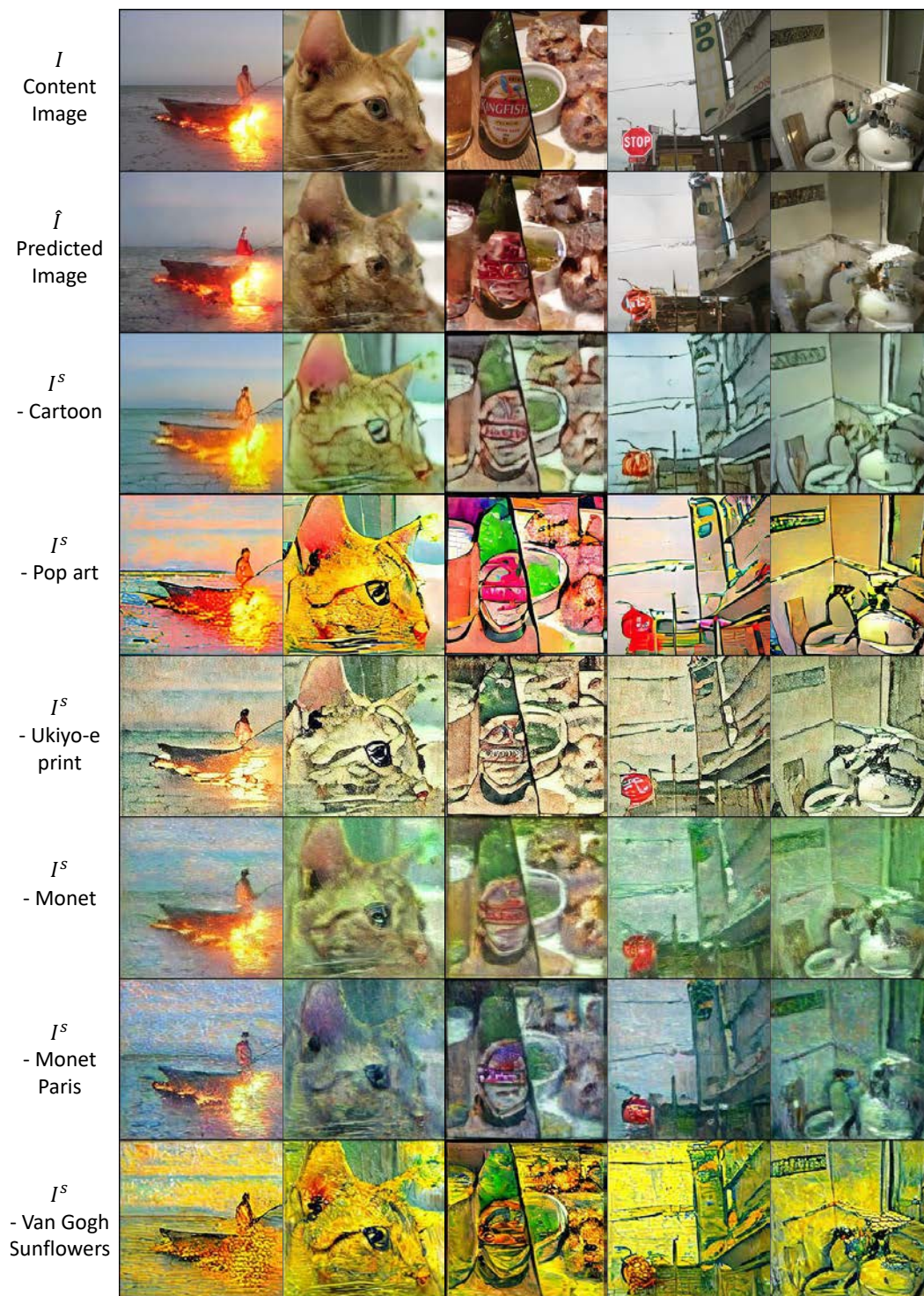


Figure 10: Additional stylized results of StylerDALLE-Ru.



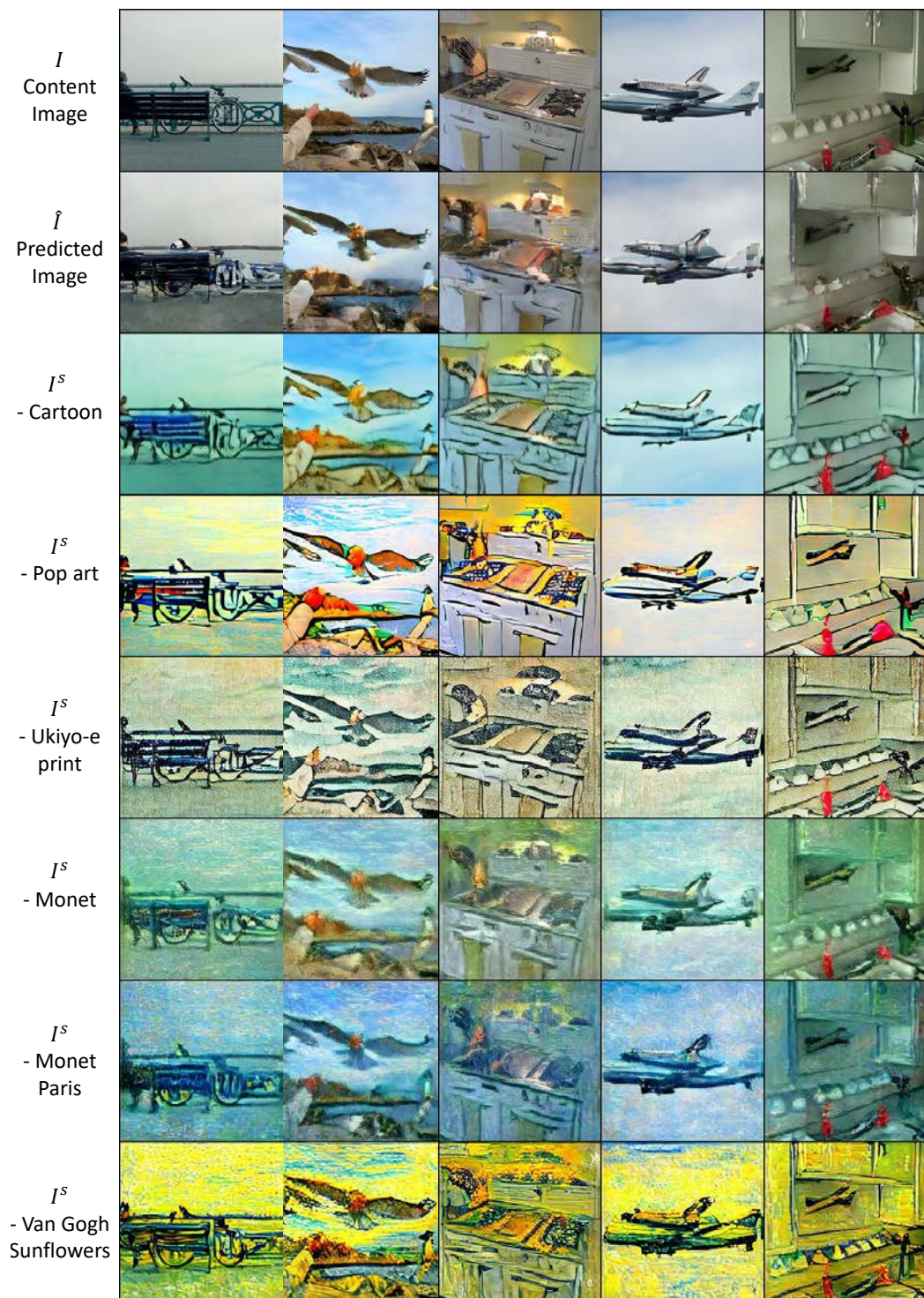


Figure 11: Additional stylized results of StylerDALLE-Ru.



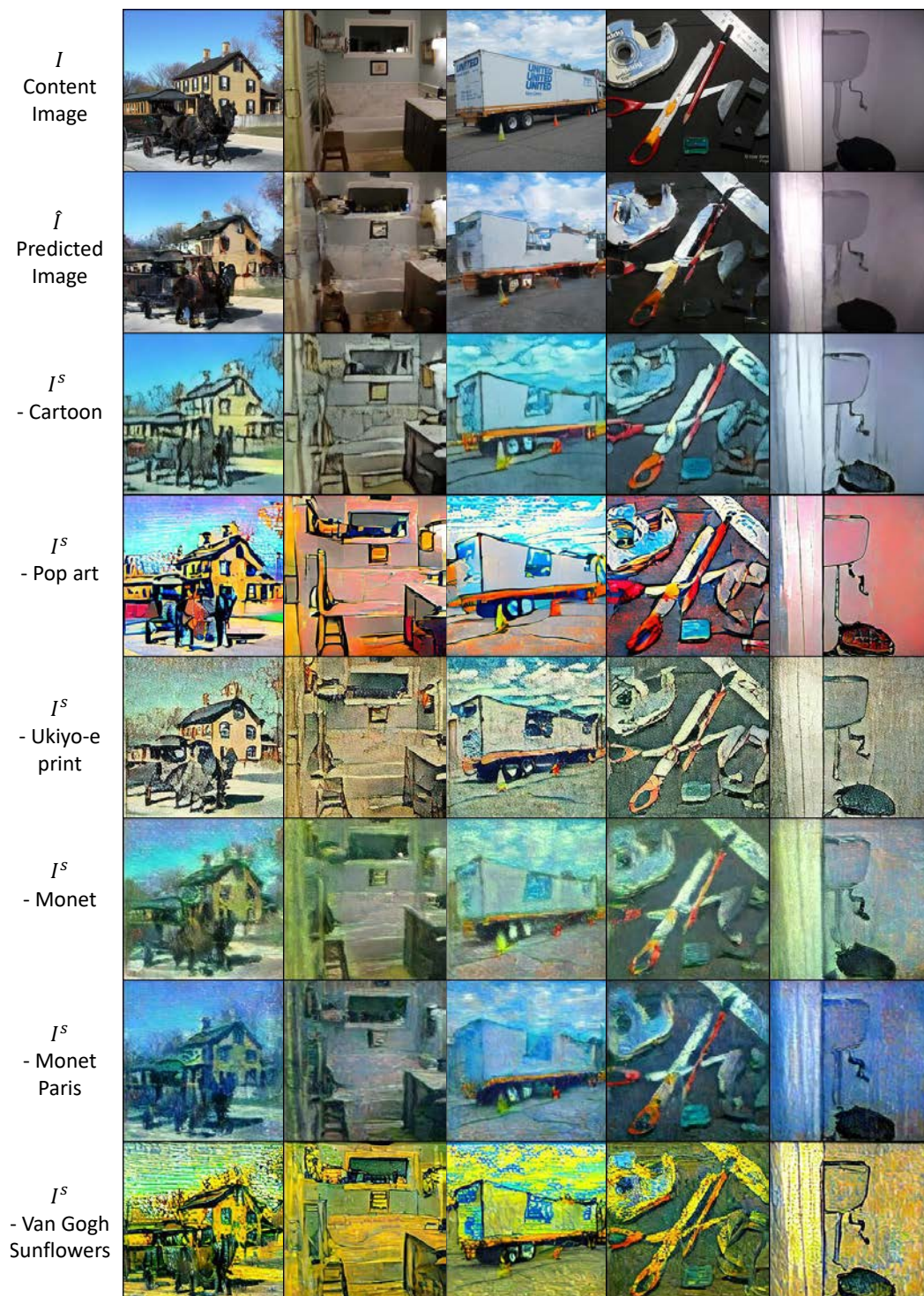


Figure 12: Additional stylized results of StylerDALLE-Ru.



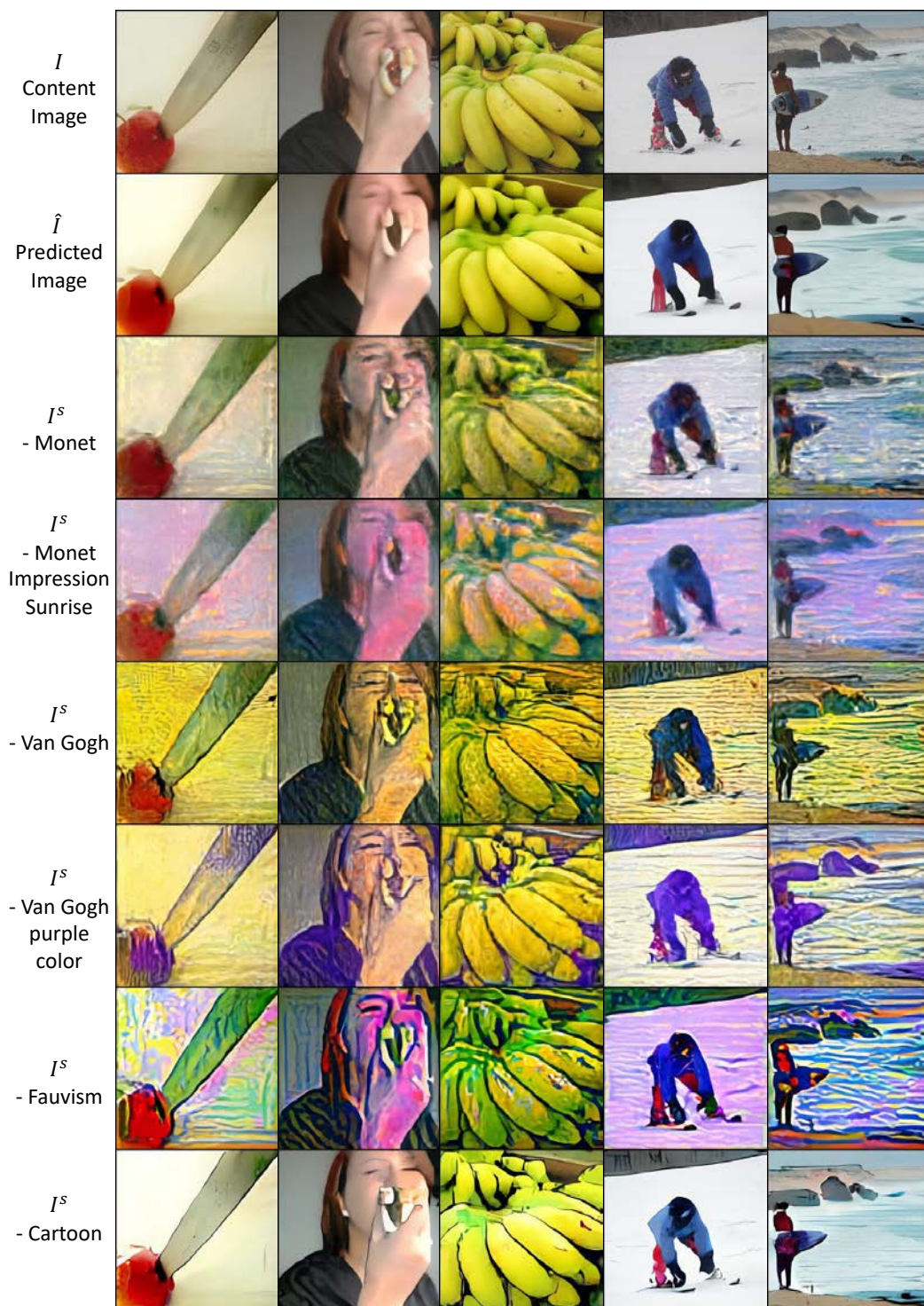


Figure 13: Additional stylized results of StylerDALLE-1.

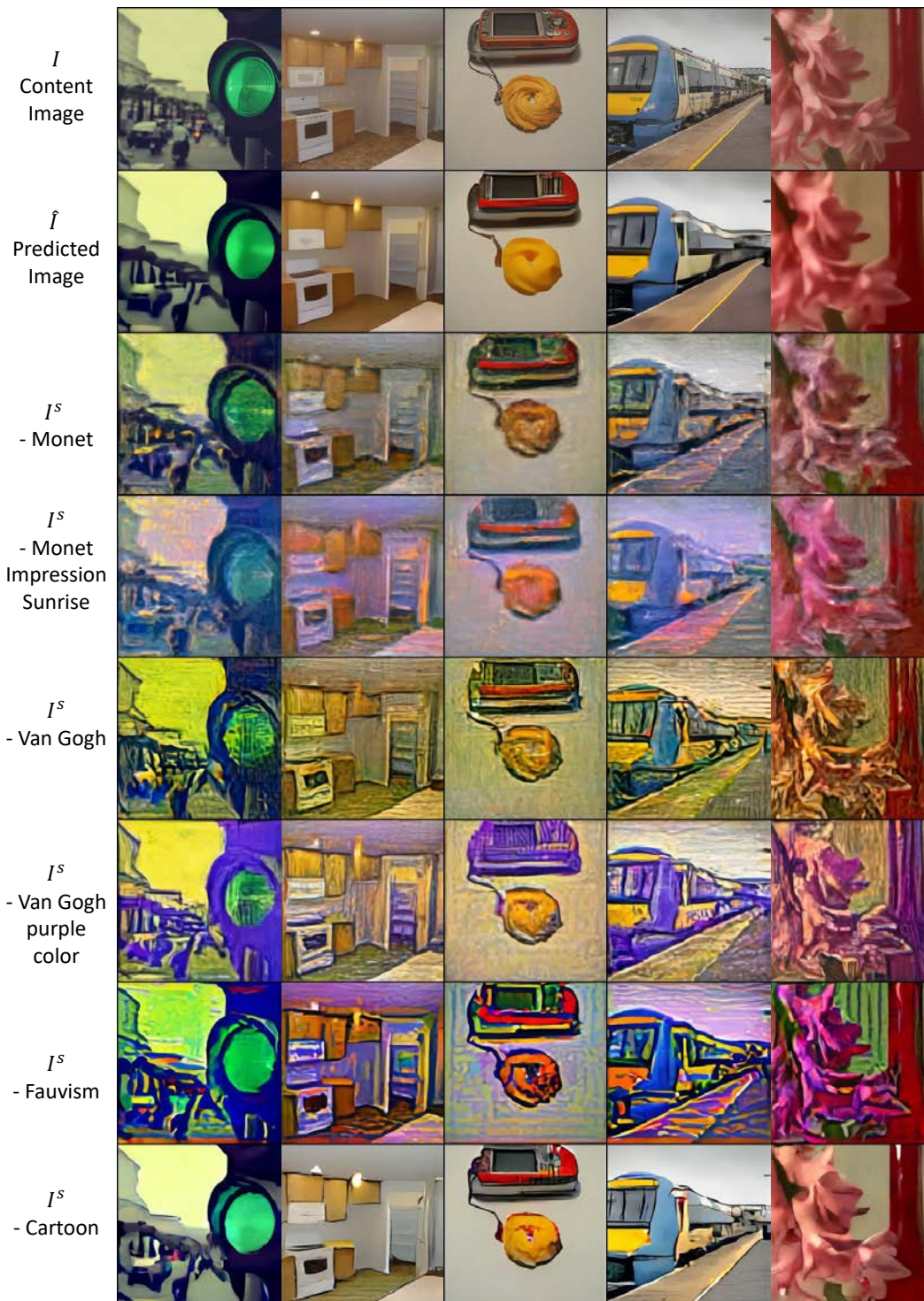


Figure 14: Additional stylized results of StylerDALLE-1.



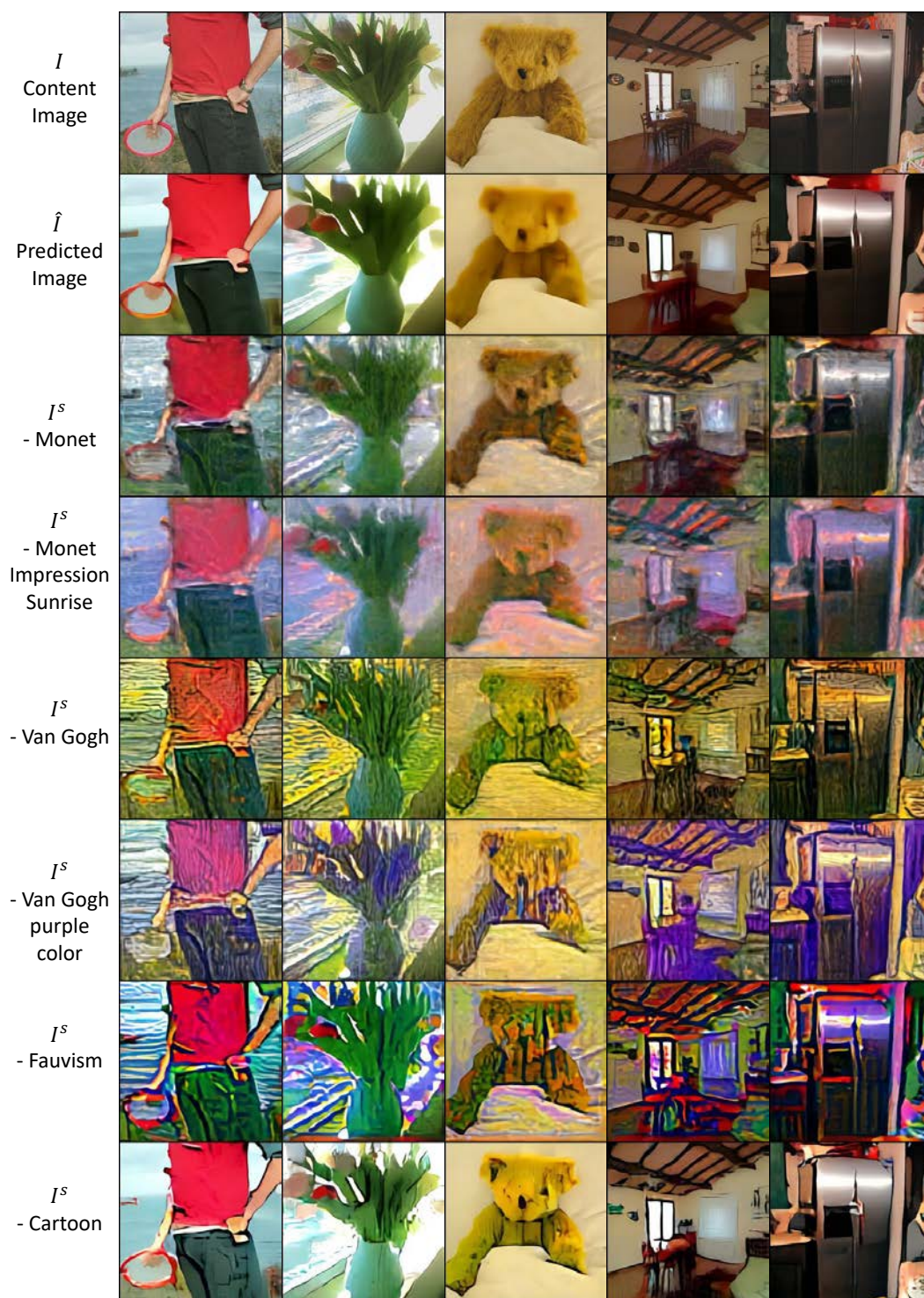


Figure 15: Additional stylized results of StylerDALLE-1.