Supplementary Material

A. Details of Implementation

A.1. Data manipulation and hyperparameters

To apply directional CLIP loss in a patch-based manner [25], we start by randomly cropping the source image. The patch size can vary, but for texture styles like golden or green crystal, we mainly use $(0.01, 0.05)$, while for artistic styles like painting by Gogh or pop art, we use $(0.01, 0.3)$. The cropped images are then augmented with perspective function and random affine transformation.

To guide both style and content, we use different weights for each loss according to the style. Although the values may differ for each style, the weights for L_{global} and L_{dir} usually range from 5000 to 30000. Additionally, Table 6 provides examples of the weights used for L_{ZeCon}, L_{VGG}, and L_{MSE}. However, one can adjust these values to improve the quality of the final image.

A.2. Patch-wise cross entropy loss for ZeCon guidance

We provide a more detailed explanation of the cross-entropy loss in equation (9). The loss takes a query patch $\boldsymbol{v}$, along with its positive counterpart $\boldsymbol{v}^+$ and N negative counterparts $\boldsymbol{v}_i^-$, where $i \in [1, \dots, N]$, as inputs. The query patch is taken from the generated image, while the positive patch is the corresponding patch from the source image. The negative patches are non-corresponding patches from the source image. The purpose of the cross-entropy loss is to encourage a patch to share the embedding space with its corresponding patch from the input, and not with the other patches. Mathematically, the cross-entropy loss can be expressed as:

$$\ell(\boldsymbol{v}, \boldsymbol{v}^+, \boldsymbol{v}^-) = -log \left[\frac{e^{\boldsymbol{v}\cdot\boldsymbol{v}^+/\tau}}{e^{\boldsymbol{v}\cdot\boldsymbol{v}^+/\tau} + \Sigma_{i=1}^{N} e^{\boldsymbol{v}\cdot\boldsymbol{v}^-{}_i/\tau}} \right] \quad (13)$$

where τ is a temperature.

A.3. Training DDIB

To compare with DDIB, we trained a diffusion model using 13,000 images from the Wikiart dataset, each with dimensions of 256×256. The model architecture is based on guided diffusion [10], with 128 base channels and attention at 16×16 and 8×8 resolutions. We did not use residual blocks for upsampling and downsampling, and we fixed the variance as a constant [16]. The model was trained for 50,000 iterations using a batch size of 8 on an NVIDIA RTX 3090.

A.4. Details of image manipulation

While patch-based CLIP losses are effective for modulating textures, they are not ideal for image translation tasks that involve translating entire classes of objects. For these tasks, guidance must be applied to the entire image rather than just small patches. Therefore, in Figures 6, 7, and 8, CLIP losses were calculated for the entire image in both image translation and manipulation tasks. However, for the image style transfer task, patch-based CLIP losses were utilized, as they are well-suited to modulating texture styles.

B. Additional experimental results

B.1. Effect of the number of timesteps

Since the diffusion process usually takes lots of time, two techniques are widely used - respacing and skipping time steps [5, 24]. The last time step T is respaced into T'. Then we forward the diffusion model to time $t_0 < T'$ and reverse the diffusion process from x_{t_0}. T' and t_0 have various effects on both image quality and time consumption. As shown in the Figure 15 (a) and (b), image quality with respect to style transformation enhances as respacing time step T' increases. However, its growth rate decreases and its difference is imperceptible even though sampling time still increases. In the mean time, CLIP score increases as t_0 increases as illustrated in Figure 15 (c). On the other hand, the content information is not fully preserved in the time steps $t_0 = 15$ or 20 as shown in the Figure 15 (a). Thus, we set (T', t_0) as $(50, 25)$ for our baseline.

B.2. Diffusion models' trade-off between style and content

With respect to style transfer, one of the challenges posed by unconditional diffusion models is to maintain content of the given image. When transforming styles of the given image, its content changes simultaneously. GAN-based methods explicitly impose content losses, such as a reconstruction loss. This results in good performance in content preservation. In contrast, diffusion models have no constraint during training phase. They generate high quality images in correspondence with the training data domain. The semantic constraints are not considered which finally results in the degradation in the quality of the generated images.

Here, we compare four diffusion models - ILVR, DDIM, DiffusionCLIP, and our proposed method - with respect to style and content in the Figure 3. ILVR utilizes downsampled reference image as condition in each reverse denoising steps. The condition helps the generated image share its content information with the reference image. However, it cannot have same identity because reverse DDPM steps without condition should be given sufficiently in order to generate images in photo style. This accordingly results in a loss of content. DDIM can reconstruct the source image when the variance of noise σ_t is set as 0. However, the style is also preserved with zero variance. When

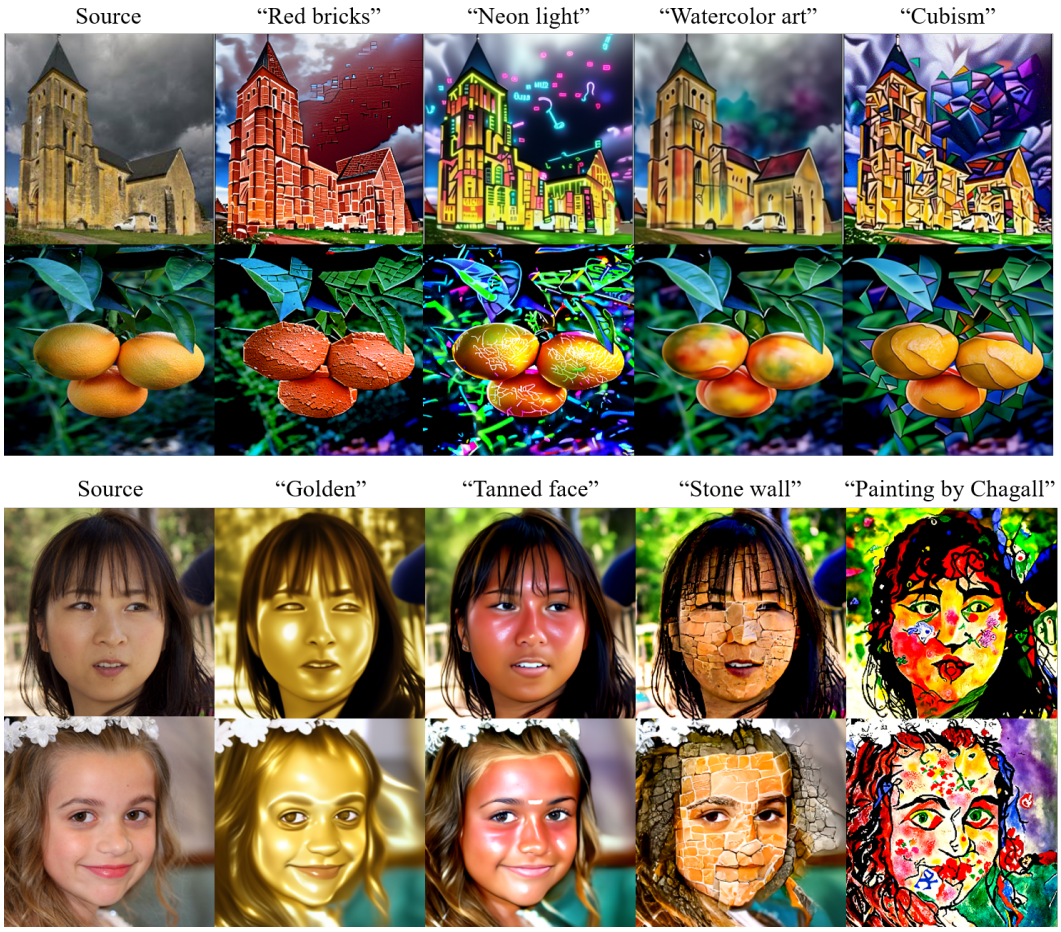

Figure 12. Additional results on various style prompts.

Figure 13. (a) Results from our baseline method. (b) Results from our method combined with timestep scheduling strategy.

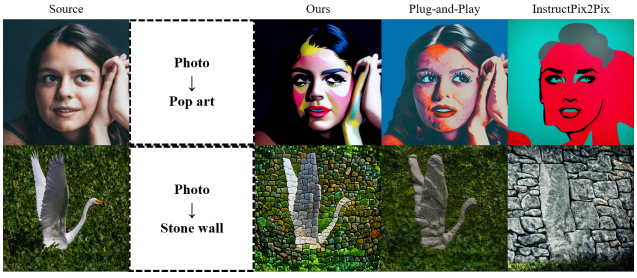

Figure 14. Comparative study results on style transfer.

we control σ_t as larger than zero, we can get photo style images but their content is altered with stochastic noise. DiffusionCLIP tried to solve the trade-off between content and style by fine-tuning the score function ϵ_θ. However, it requires much more time due to the model training for each style and data preparation. In addition, the content cannot be maintained when the source images are from unseen domains. In contrast, our proposed loss $\mathcal{L}_{\text{ZeCon}}$ does not require additional networks or fine-tuning. This leads to shorter time than DiffusionCLIP. With the help of ZeCon guidance, we could retain the content of source image from any domain and translate it into different styles.

B.3. More Comparison with GAN and CNN-based methods

In addition to the comparative studies on the GAN-based methods in the Section 4.2, we conducted more comparisons with various GAN-based and CNN-based methods including both text and image guidance. For text guidance method, we compared our method with one more method, LDAST [13]. For image guidance, we included three methods, SANet [12], AdaIN [18], and WCT2 [39]. As shown in the Figure 16, we could notice that LDAST and WCT2

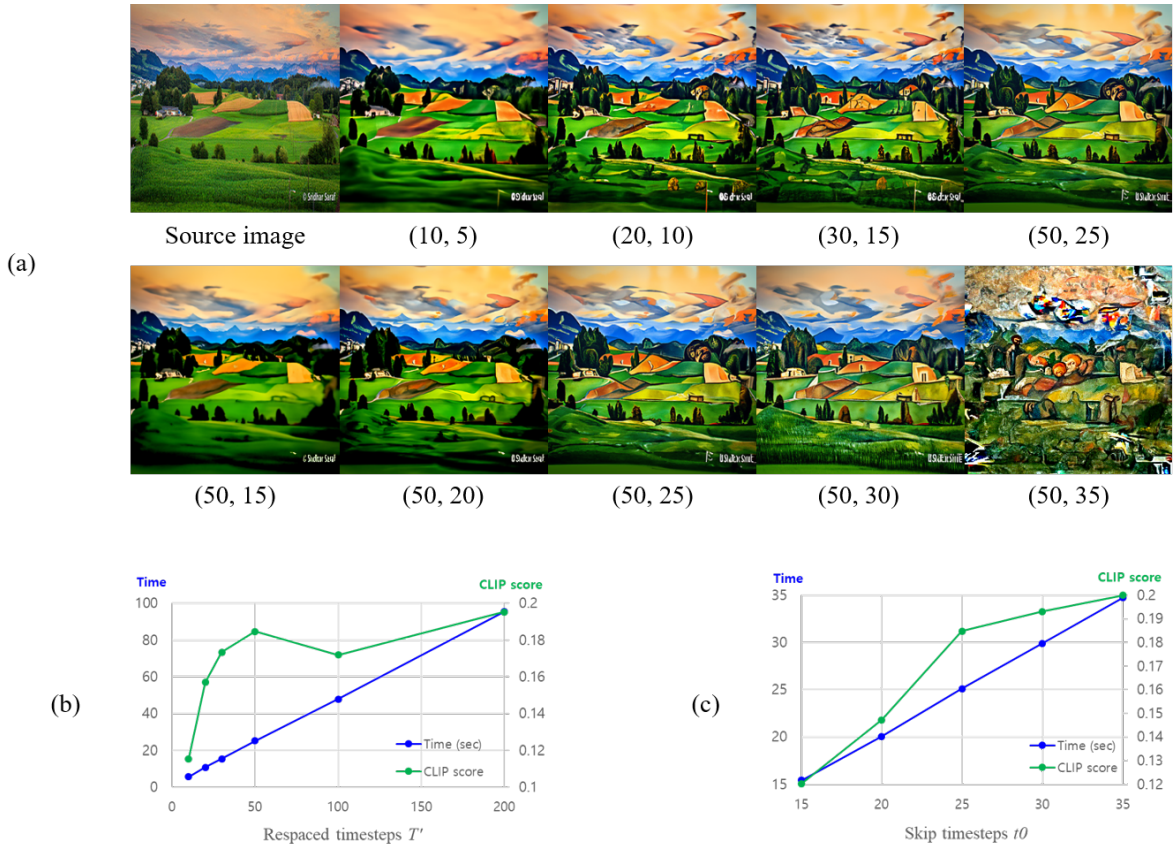

Figure 15. The effect of the respaced time T' and skipped time t_0. (a) demonstrates the images sampled with (T',t_0). The first row shows the difference between various T' when t_0 is its half and the second row shows the difference between various t_0 when $T' = 50$. (b) and (c) illustrates the relationship between sampling time and CLIP score as graphs for the first and second rows of (a), respectively.

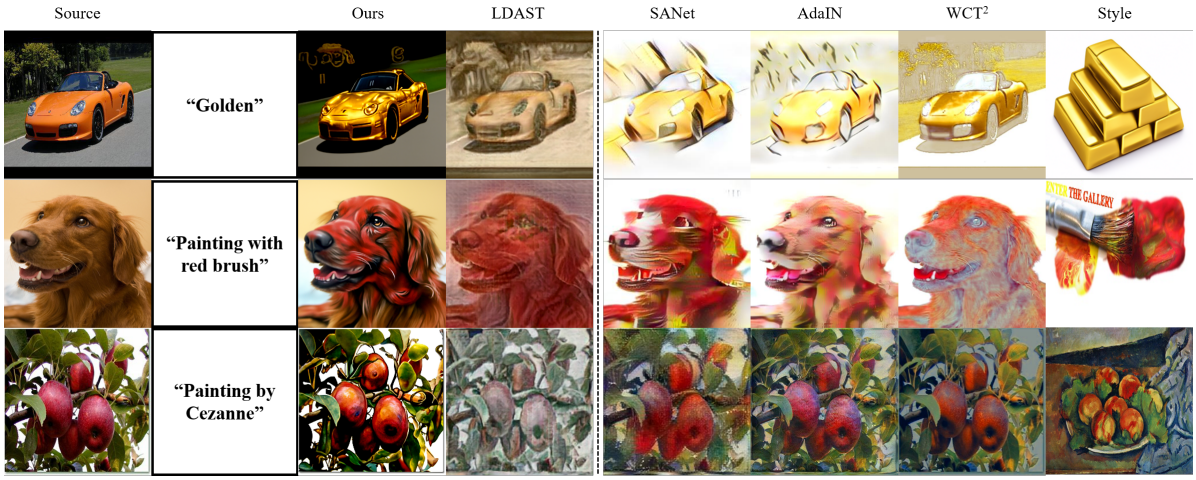

Figure 16. Comparative study results.

could preserve the content information better than SANet and AdaIN. However, all the four methods for comparison show inferior performance than our method in perspective of style transformation.

B.4. More Comparison with Diffusion-based methods

To extend our comparative studies for style transfer, we compared with both Plug-and-Play[38] and

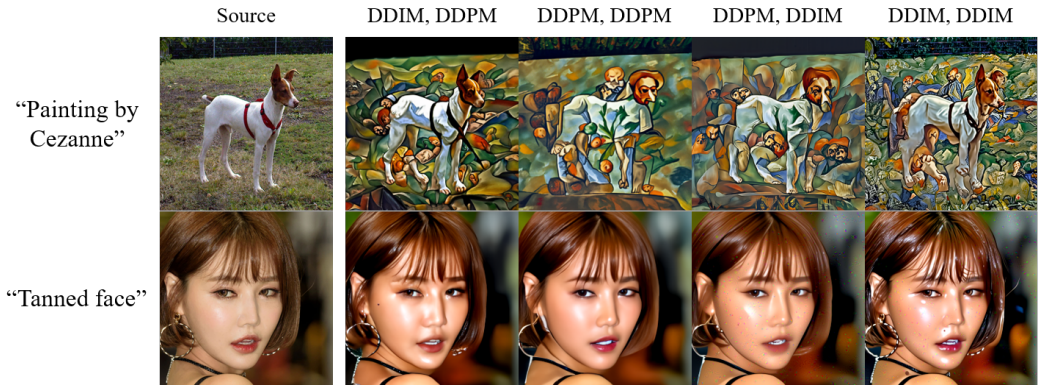

Figure 17. Ablation study results on diffusion processes. From the second column to the right, the combinations of methods (forward, reverse) are (DDIM, DDPM), (DDPM, DDPM), (DDPM, DDIM), and (DDIM, DDIM).

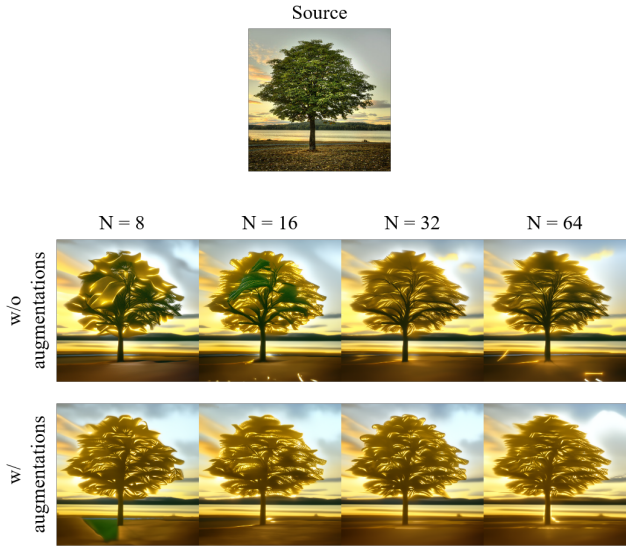

Figure 18. Ablation study results on augmentation and the number of patches N. The target prompt is “Golden.”

InstructPix2Pix[3] As can be seen in Figure 14, InstructPix2Pix struggled with content preservation and Plug-and-Play fell short in accurately translating into the target style. Contrarily, our proposed method showed outstanding performance in preserving content and accurately translating style.

B.5. Different timesteps on content and style loss

We orchestrated content and style loss applications at different timesteps. Initially, only content losses were imposed during the earlier timesteps. Subsequently, style losses were exclusively applied later in the process. As depicted in Figure 13, outcomes using the timestep scheduling strategy excel at preserving detailed content, like eyebrows and glasses. Nevertheless, the texture transformation, specifically into wood, is not sufficiently pronounced in these results.

B.6. DDPM and DDIM for diffusion processes

Although either DDPM or DDIM can be utilized for both forward and reverse processes, we conducted a comparative study in order to show their differences in the generated images. As shown in Figure 17, results from the forward DDIM show better performance in preserving content than DDPM. The earring in the second row appears unchanged in the output images generated by forward DDIM, whereas its shape is altered in the images produced by forward DDPM. For reverse process, DDPM tends to transform styles better compared to DDIM. Accordingly, we chose to use DDIM as forward and DDPM as reverse process as default.

B.7. Unseen domains

DiffusionCLIP tried to solve the trade-off between content and style by fine-tuning the diffusion model with identity loss. Because of the constraints imposed on the fine-tuned model, the transformed image shows high performance in identity preservation. However, the fine-tuned model $\hat{\epsilon}_\theta$ converts only the photo domain images. When it comes to unseen domains, such as portraits or paintings, they should be converted to photo images through ϵ_θ. Since it has not been fine-tuned with identity loss, the semantic information is lost during the reverse sampling process due to the stochastic property of the diffusion model. Thus, the final output from images of unseen domains is degraded in its quality. In contrast, our proposed method can transform even the unseen domain images with only one step. Since our method can preserve the identity with content guidance, the final outputs do not suffer from quality degradation. Also, our method takes about 38 seconds whereas DiffusionCLIP requires about 400 seconds for model fine-tuning and sampling. As described in the Figure 3, DiffusionCLIP requires two steps from portrait to photo to Pixar domains. In this process, the face identity is destroyed. However, the

Model	Style prompt	CLIP-global	CLIP-directional	ZeCon	MSE	VGG	Patch size	t_0
ImageNET	Golden	5000	5000	100	5000	10	0.05	15
	Watercolor art	5000	10000	300	0	100	0.3	25
	Stained glasses	15000	15000	200	1000	10	0.05	25
	Oil painting of flowers	20000	20000	1500	10000	10	0.05	25
	Red bricks	20000	40000	1000	1000	10	0.05	25
	Wooden	20000	50000	1000	1000	10	0.05	25
	Leather	20000	30000	2000	20000	200	0.3	25
	Marbling	20000	30000	2000	20000	200	0.3	25
	Autumn	20000	20000	700	10000	100	0.05	25
	Snowy	20000	20000	700	0	100	0.05	25
FFHQ	Pop art	10000	20000	50	1000	50	0.3	25
	Stone wall	2000	50000	500	5000	10	0.1	25
	Tanned face	15000	15000	1000	10000	100	0.3	25
	Clay	40000	40000	1000	10000	0	0.05	25
	Portrait by Gogh	10000	7000	10	3000	50	0.3	25
	A sketch with crayon	10000	20000	500	10000	100	0.3	25
	3d render in the style of Pixar	5000	5000	500	10000	100	0.3	25
	Golden	7000	7000	200	0	50	0.05	15
	Ukiyo-e	8000	20000	1000	5000	100	0.3	25
	Marbling	20000	40000	1000	10000	10	0.3	25

Table 6. Examples of hyperparameters for various style prompts. Weights for CLIP-global loss, CLIP-directional loss, ZeCon loss, MSE loss, and VGG loss are given. For patch-based CLIP guidance, we control the patch size. The maximum size is given in the table with the minimum of 0.01. t_0 is the time step to which the source image is forwarded when $T' = 50$.

Figure 19. Generated images using random seeds. The style prompts of the first and the second rows are "Leather" and "A sketch of crayon", respectively

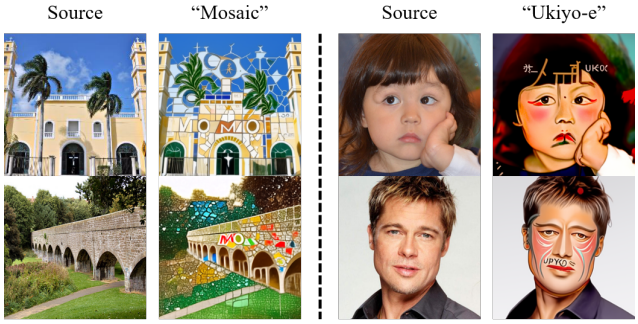

Figure 20. Failure cases. Texts from target prompts sometimes appear on the generated images.

proposed method could preserve the identity while transforming into the style of Pixar.

B.8. Augmentation of patches

We use patch-based CLIP losses for style guidance. From the denoised image $\hat{x}_{0,t}$, N patches are randomly cropped and augmented with perspective function and random affine transformation. In order to check the importance of augmentation, we conducted ablation study on the augmentation and the number of patches. As shown in the Figure 18, the images generated without augmentation could not transform to gold enough. Also, when $N < 32$, the tree is not sufficiently converted to align with the target prompt "Golden". From these observations, we chose $N = 32$ with augmentations and this choice results in further reduction in inference time to 24 seconds.

B.9. Stochastic property

The random nature of DDPM leads to various modifications generated from the same style prompt. As shown in Figure 19, we observed that the same image and text prompt pair could generate various images using different random seeds.

B.10. User study

For quantitative analysis, we conducted a user study. For comparison with GAN-based methods, 60 images with four styles have been used in total. The styles involved are "golden", "clay", "3d render in the style of Pixar", and "pop art." We utilized human face images because StyleCLIP and StyleGAN-NADA are based on face dataset. In addition, we totally generated 24 images with six styles for comparison with DiffusionCLIP. We chose three styles ("neon light", "green crystal", and "Ukiyo-e") for the photo domain and the other three styles ("3d render in the style of Pixar", "pop art", and "golden") for unseen domains. We used portraits and paintings from Wikiart dataset for unseen domain images. Besides, for ablation study on content losses, we used 15 images for three styles ("golden", "oil painting of flowers", "leather") in total. The number of questions were 14. 21 users participated in the user study. They were randomly recruited online.

C. Limitations

Although our proposed method has various strengths and shows great performance, there remain some limitations. As described in the Appendix A.1, one should find weights for each loss though their relevant ranges are given in this paper. Also, it has been observed that in some cases, the text prompts describing the targeted style are displayed on the generated images (Figure 20).

Figure 21. Additional results on various style prompts.

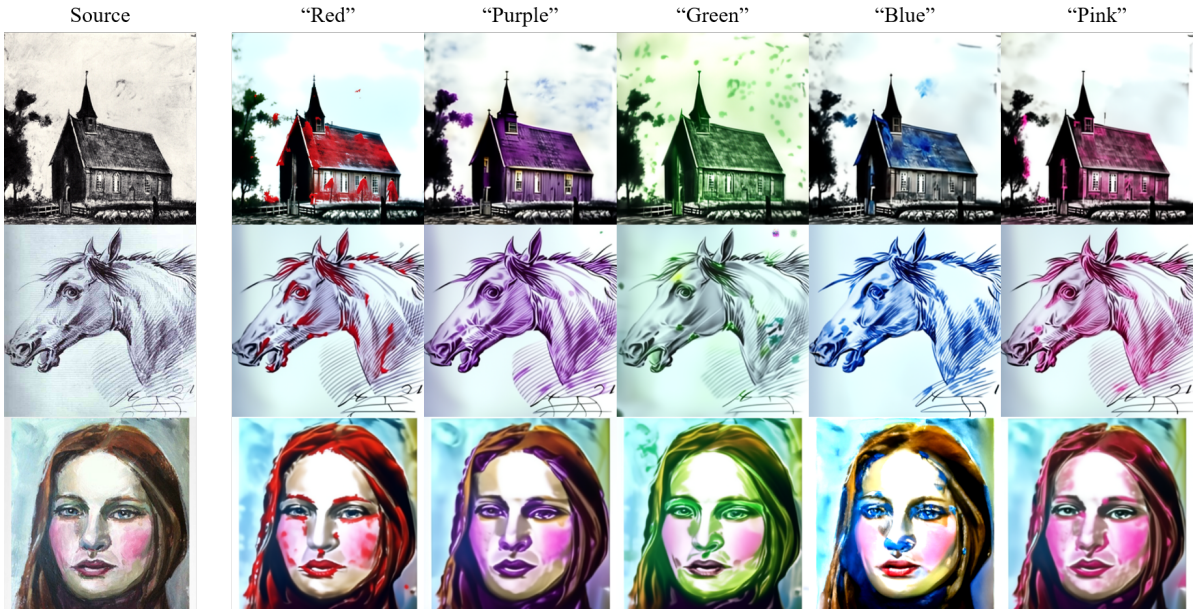

Figure 22. Additional results on color style prompts.

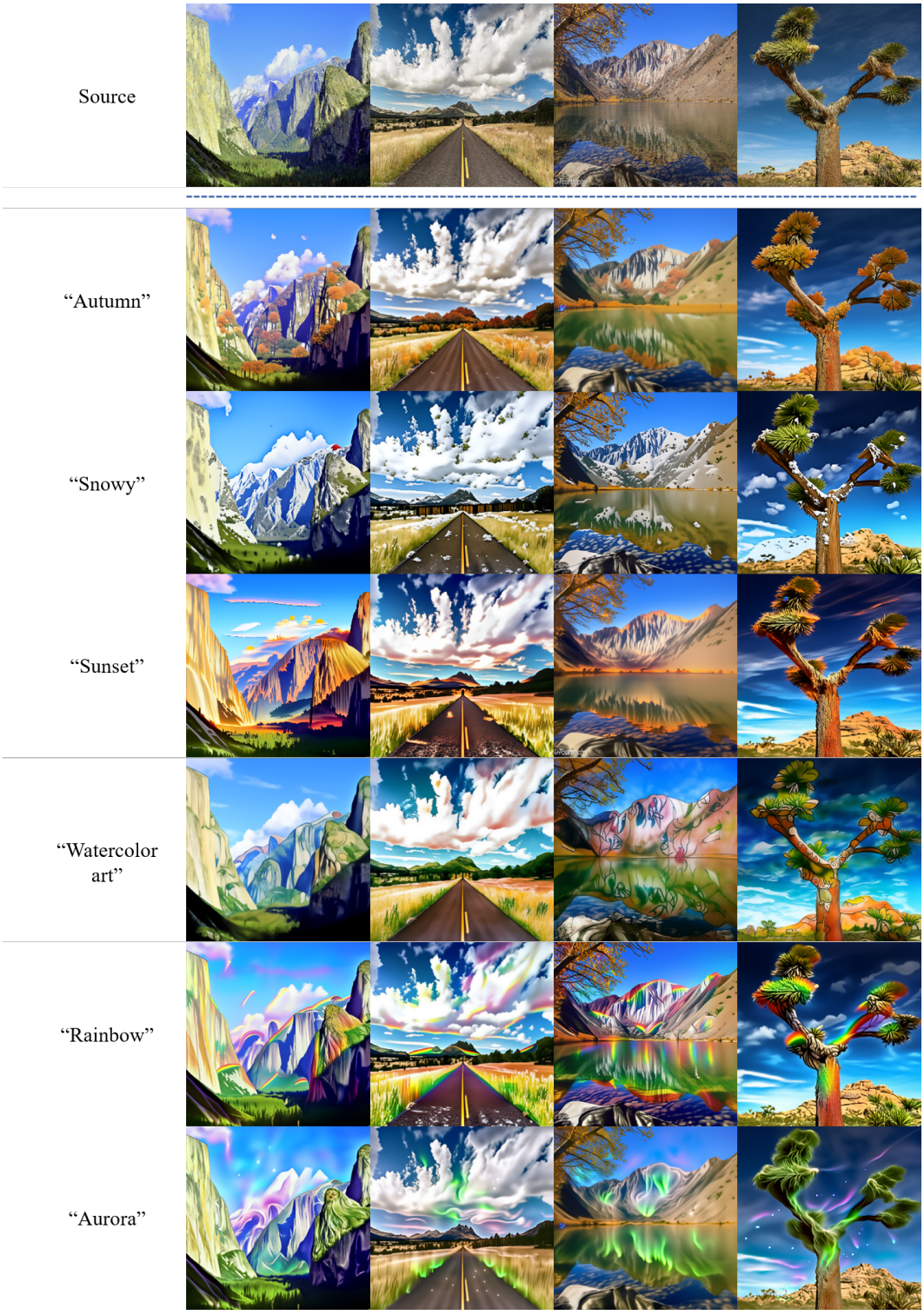

Figure 23. Additional results on various style prompts.

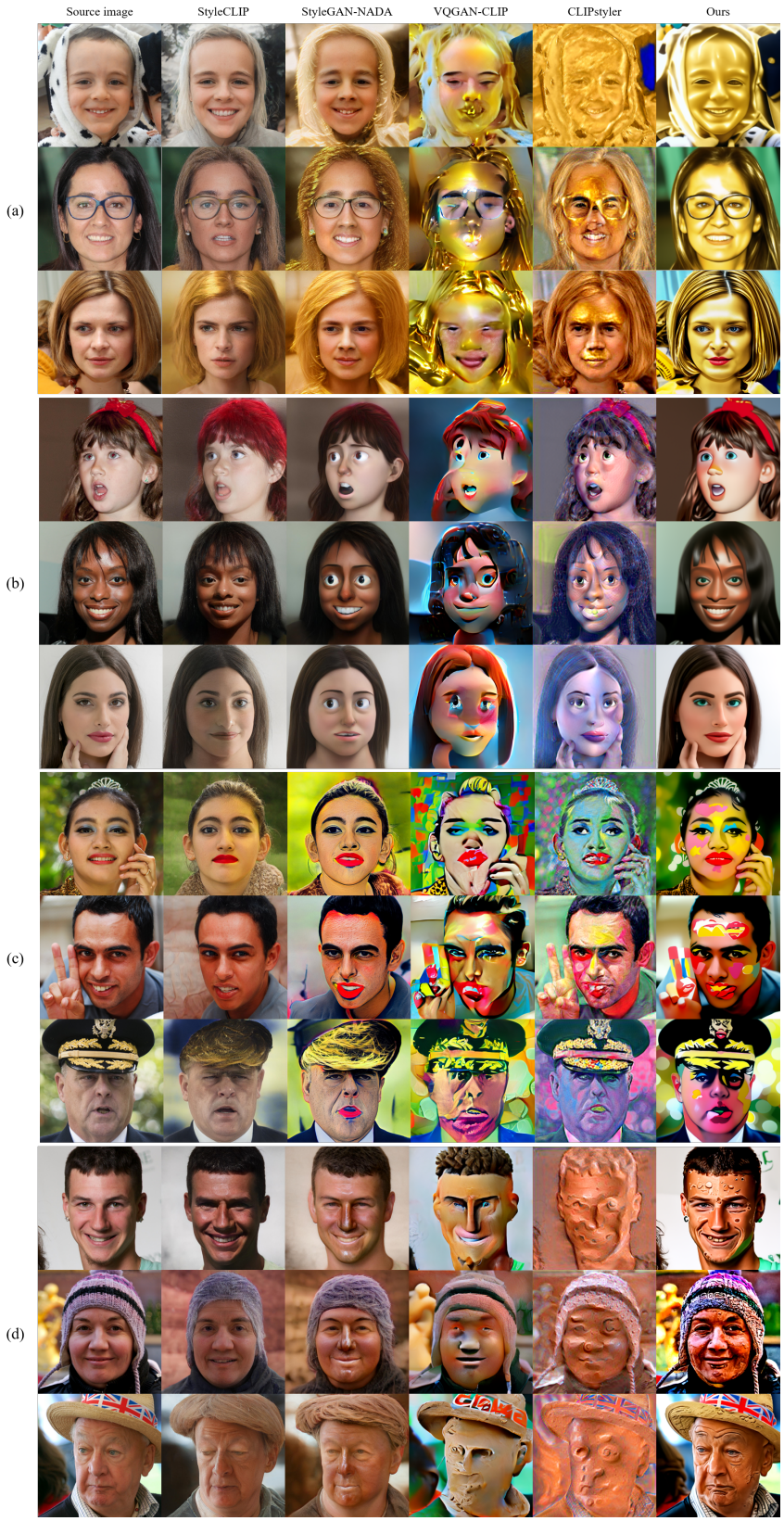

Figure 24. Additional results on the comparative studies. (a), (b), (c), and (d) are the results on styles of “golden”, “3d render in the style of Pixar”, “pop art”, and “clay”, respectively.

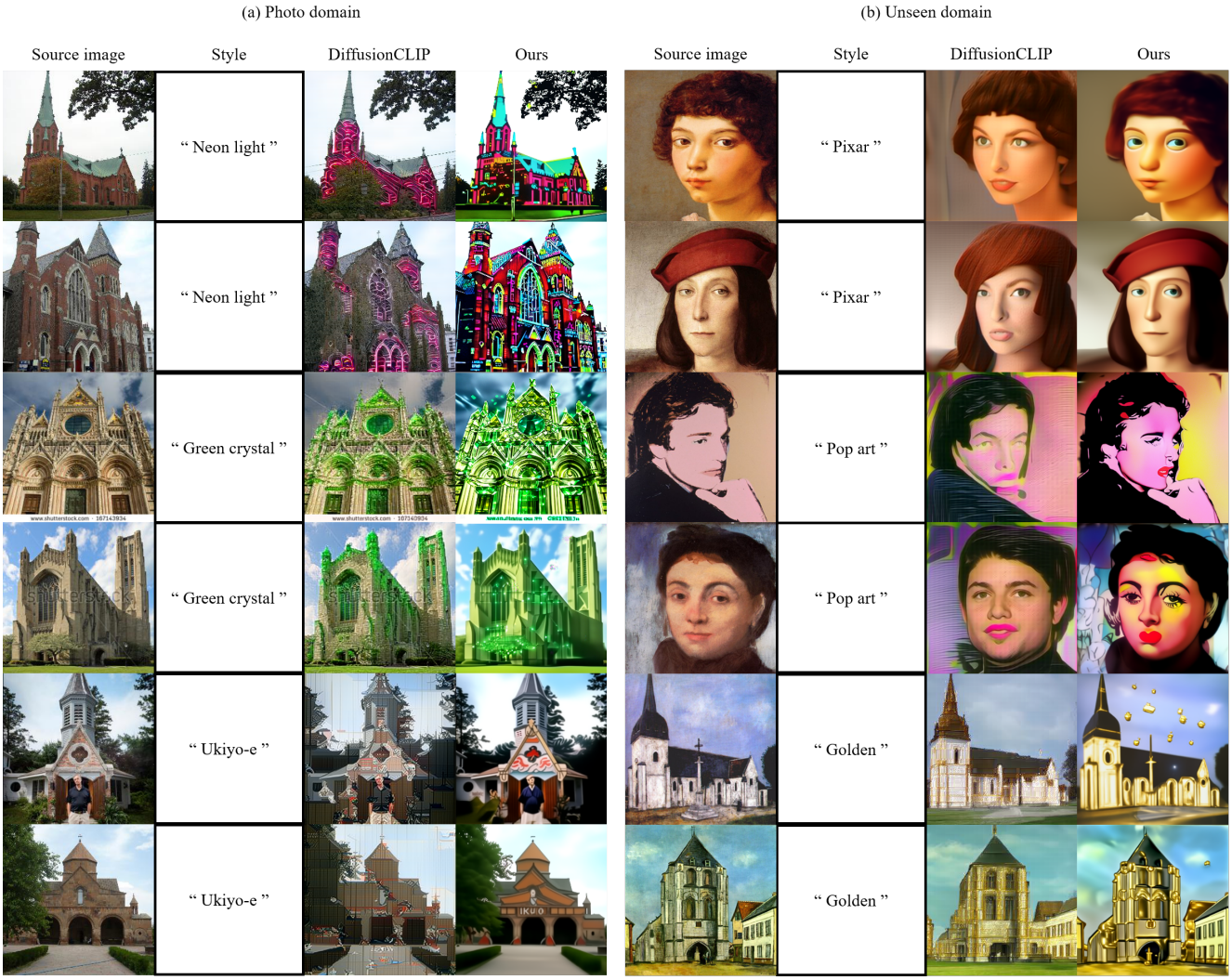

Figure 25. Additional results on the comparative studies. Source images in (a) are from photo domain and ones in (b) are from unseen domains such as portrait or painting.