

StyleStudio: Text-Driven Style Transfer with Selective Control of Style Elements

Supplementary Material

This supplementary materials provide additional details regarding the experimental setup described in the main paper and offer an extended analysis of the contributions of individual components. The content is organized as follows:

- **Details of Experiments.** This section provides additional information about the experiments discussed in the main paper, including specifics on the quantitative evaluations and the user study setup.
- **Ablation Study.** Qualitative comparisons from the ablation experiments are presented, analyzing the impact of the Teacher Model, particularly in terms of timestep selection and the choice of Attention Map.
- **Additional Qualitative Comparisons.** This section presents extensive qualitative comparisons, demonstrating that cross-modal AdaIN effectively prevents style overfitting, while the Teacher Model ensures layout stability and mitigates the occurrence of artifacts.
- **Integration with Other Methods.** This section explores how our approach can be integrated with existing methods, such as InstantStyle [34] and StyleCrafter [19], showcasing its ability to enhance their performance and adaptability.

A. Implementation Details

We set the random seed to 42 for reproducibility, used 50 inference steps, and applied a uniform guidance scale of 5 across all methods. In the qualitative and quantitative comparison experiments, for the implementation involving the Teacher Model, its participation was specifically limited to the first 20 steps. All experiments were conducted on a single NVIDIA GTX-4090 GPU.

Adapter-based methods [9, 34, 37, 38] are particularly suitable for style transfer. Their fine-tuning-free nature, combined with high-quality style transfer performance, has made them widely adopted. CSGO [37] employs a widely used adapter-based model structure and is the first method trained on a meticulously curated dataset specifically designed for style transfer. This effectively decouples the content and style in style images, enhancing the grasp of style details such as brushstrokes and textures. Therefore, in the experimental section, we selected it as the baseline and implemented specific modifications based on it. The implementation details are as follows:

We only retained the modules in CSGO [37] related to text-driven style transfer, removing irrelevant components, *e.g.*, ControlNet [39]. This optimization reduces potential interference while lowering experimental costs, including memory usage during inference. At the same time, both the

Teacher Model and cross-modal AdaIN are optional and can be used based on specific needs. For the quantitative experiments in the main paper, we incorporated both the Teacher Model and the cross-modal AdaIN module to achieve optimal text alignment. In the qualitative and quantitative comparison experiments, the Teacher Model participated for the first 20 time steps, with the total number of inference steps set to 50.

Algorithm 1 SDXL-Guided Self-Attention Replacement

Input: P_{dst} : a target prompt; I_{ref} : style reference image; S : random seed; DM: raw Stable Diffusion Model; ST: style transfer Method Model; t_{cutoff} : stop replacement time step;
Output: I_{style} : text-driven stylized image;
1: $z_T \sim \mathcal{N}(0, 1)$, a unit Gaussian random value sampled with random seed S ;
2: $z_T^* \leftarrow z_T$;
3: **for** $t = T, T - 1, \dots, 1$ **do**
4: **if** $t > t_{cutoff}$ **then**
5: $z_{t-1}, M_{self} \leftarrow \text{DM}(z_t, P_{dst}, t)$;
6: $z_{t-1}^* \leftarrow \text{ST}(z_t^*, I_{ref}, P_{dst}, t) \{M_{self}^* \leftarrow M_{self}\}$;
7: **else**
8: $z_{t-1}^* \leftarrow \text{ST}(z_t^*, I_{ref}, P_{dst}, t)$
9: **end if**
10: **end for**
11: **Return** $I_{res} \leftarrow \text{Decoder}(z_0)$;

B. Evaluation Settings and User Study

In the quantitative experiments presented in the main paper, the evaluation was conducted using prompts derived from StyleAdapter [35], with specific examples provided in Fig. 12. The style images were randomly sampled from the test set of StyleShot [9], with representative examples shown in Fig. 13. Ultimately, each method generated 1,000 images for the quantitative experiments.

Beyond quantitative evaluations, we conducted a user study to gain subjective insights into the performance of different methods. The study involved 12 pairs of reference images and prompts. For each pair, participants were asked to assess and select the method they found superior based on two criteria: text alignment and style similarity. To ensure a fair assessment, participants were provided with a brief explanation of the task and evaluation criteria beforehand. We collected responses from 49 participants with diverse backgrounds, including individuals with relevant expertise in text-to-image tasks. The specific design of the questionnaire, including example pairs and evaluation guidelines, is shown in Fig. 16.

<p>A robot. A girl wearing a red dress, she is dancing. A boy wearing glasses, he is reading a thick book. a little cute boy. A woman wearing a green sportswear, she is running. A woman wearing a purple hat and a yellow scarf. A man wearing a black leather jacket and a red tie. A little boy with glasses and a watch. A smiling little girl. A little boy playing football. An curly-haired boy. A little girl holding flowers. A lovely kitten walking in a garden. A puppy sitting on a sofa. A fluffy white rabbit with pink ears and nose. A brown puppy with black spots and a red collar. A black and white panda. A dog in a bucket. A cat wearing a hat. A cute little fish in aquarium. A bird in a word. A kitten sleeping on a pillow. A parrot singing a song. A monkey playing with a banana. A turtle wearing sunglasses. A hamster eating a carrot.</p>	<p>A white rose. A sunflower smiling at the sun. A cactus wearing a hat. A daisy with a ladybug on it. A pine tree with a snowman hugging it. A mushroom in winter. A beautiful lotus. A lotus with a frog meditating on it. A cherry blossom. A palm tree. A river with rapids and rocks. A creek with clear water and colorful pebbles. A lake with calm water and reflections. A waterfall with mist and rainbows. A stone with a face carved on it, standing on a pedestal in a museum. A stone with a hole in it. A stone with a pattern of stripes on it. A stone with a crack in it, holding a plant growing out of it. A snowy mountain peak. A mountain goat on a cliff. A red baseball cap. A football on the grass. A motorcycle. A modern house with a pool. A house made of cardboard boxes. A house covered with ice and snow.</p>
--	--

Figure 12. Details of the Test Set. The prompts used in the quantitative experiments were derived from StyleAdapter [35].

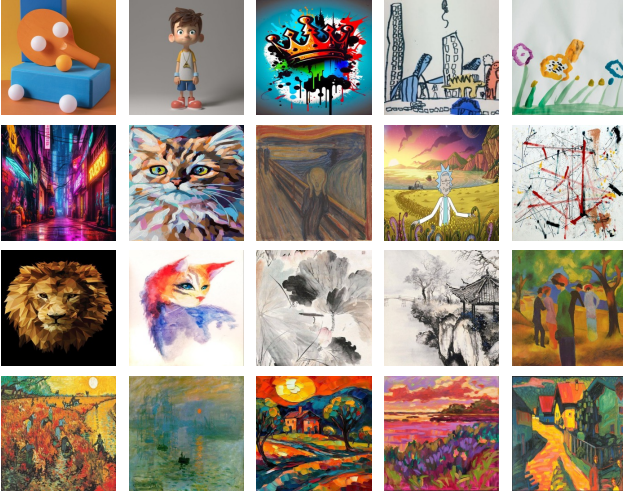


Figure 13. Details of the Test Set. The style images used in the quantitative experiments were randomly sampled from the test set of StyleShot [9].

C. Additional Ablation Study

Qualitative Results of the Ablation Study. While the main paper presents a quantitative analysis, a qualitative comparison provides a more intuitive understanding of the contributions of each component. By incrementally integrating the corresponding components, we demonstrate their individual effects. Fig. 17 showcases representative visual outcomes from our qualitative experiments. A comparison between the second and third columns highlights that cross-modal AdaIN significantly improves text alignment while preserving style similarity. Furthermore, as shown in the green apple example, introducing the Teacher Model not only enhances layout stability but also resolves remaining artifacts, ensuring spatial consistency across different styles.

Self-Attention map and layout stability. In the UNet of Stable Diffusion [23, 27], Cross-Attention [33] primarily aligns the prompt with the generated image, determining how textual input influences the overall style and content. Self-Attention [33], on the other hand, focuses on the in-

ternal coherence of the image, maintaining spatial relationships and structural consistency. As shown in Fig. 18, swapping the Self-Attention Map ensures layout stability and consistency across different styles of images, whereas replacing the Cross-Attention Map fails to achieve this effect, resulting in noticeable differences in the main layout under varying styles. All experiments were conducted by adding the Teacher Model to the baseline CSGO framework. To objectively evaluate the impact of the Teacher Model, cross-modal AdaIN was not used in these experiments, isolating the Teacher Model’s contribution to layout stability.

Choice of Teacher Model participation timestep. To evaluate the impact of the Teacher Model’s participation timestep on the final generation results, we conducted experiments analyzing its effect. The Teacher Model is designed to ensure layout stability while avoiding artifacts, such as checkerboard patterns. To objectively evaluate the impact of the Teacher Model, cross-modal AdaIN was not used in these experiments. As shown in Fig. 19, the term “timestep” refers to the number of denoising steps during which the Teacher Model is active. The results demonstrate that insufficient participation (short timesteps) fails to resolve layout issues, while prolonged involvement (long timesteps) negatively affects the final style fidelity. Rows 3 and 4 illustrate that even small changes in the timestep significantly influence the results, while Rows 5 and 6 show that the optimal timestep can vary across different styles. Based on these findings, a timestep between 10 and 20 strikes a reasonable balance between layout stability and style preservation.

Compare with image-based style transfer(I2I). Although our method utilizes the Self-Attention Map provided by the Teacher Model, this does not equate to I2I. As shown in Fig. 20, the I2I approach provided by CSGO [37] fails to preserve the color information of the content image effectively. In contrast, our method can more accurately adhere to the prompt’s description. To ensure fairness, the noise used in our method is identical to that used in generating the content image.

D. Additional Comparisons

Qualitative experiments are conducted to visually demonstrate the strengths of our method, particularly in capturing style details and ensuring alignment with the given textual descriptions. This allows for a more intuitive comparison with state-of-the-art methods, showcasing the superior performance of our approach in real-world scenarios. We provided additional qualitative comparisons between our method and state-of-the-art approaches to better illustrate the strengths and weaknesses of each method.

In Fig. 24, our method outperforms others in both overall style similarity and the ability to capture fine de-



















Style		Ours	CSGO	InstantStyle	IP-Adapter	StyleCrafter	StyleAlign	DEADiff	StyleShot
	<i>"A snowy mountain peak."</i>								
	clip style similarity	0.642	0.661	0.639	0.650	0.732	0.720	0.606	0.659
	dino style similarity	0.467	0.289	0.395	0.350	0.721	0.723	0.315	0.525
	<i>"A cute little fish in aquarium."</i>								
	clip style similarity	0.530	0.629	0.672	0.829	0.802	0.867	0.667	0.691
	dino style similarity	0.184	0.216	0.194	0.433	0.584	0.711	0.311	0.396

Figure 14. We observed that existing metrics generally fail to capture adherence to style. They tend to favor higher semantic similarity to the style image rather than better style transfer, a known issue often referred to as content leakage. A higher semantic similarity score does not indicate better style preservation and can, in fact, weaken the style in the generated results.



Figure 15. More results of Style-Based CFG.

tails, such as textures. Additionally, it achieves the highest accuracy in aligning with the prompt descriptions. For methods based on the Stable Diffusion XL [23], approaches like CSGO [37] and InstantStyle [34] exhibit noticeable style overfitting, while IP-Adapter [38] and StyleCrafter [19] tend to suffer from content leakage. Meanwhile, StyleAlign [36] produces results of relatively lower quality. For methods based on the Stable Diffusion 1.5, DEADiff [24] struggles with accurately capturing the style, and although StyleShot [9] performs reasonably well in capturing style, it still encounters issues such as content leakage. Content leakage can indeed be seen as a form of overfitting to the style reference, where the model overly relies on the style image, causing elements of the style reference to dominate or intrude on the content representation. This highlights a lack of proper disentanglement between style and content in such cases.

A more nuanced form of style overfitting, as discussed in this paper, arises when text-driven style transfer methods struggle to adapt to nuanced variations in prompt de-

tails, such as changes in color. The challenge lies in whether these methods can accurately align with the evolving prompt descriptions while preserving the integrity of the style. This aspect is further validated in Fig. 25. Methods such as CSGO [37], InstantStyle [34], and StyleShot [9] struggle to differentiate the color specifications described in the prompt. Additionally, IP-Adapter [38] and DEADiff [24] face challenges with style dissimilarity, while StyleCrafter [19] demonstrates some bias toward the structure of the style reference, particularly evident in the “car” example.

In the main paper, we also focus on the issue of layout stability. Through extensive experiments, as shown in Fig. 26, we demonstrate that our method can effectively ensure layout stability. CSGO [37] frequently exhibits artifacts such as checkerboard patterns, while other methods also encounter issues with layout instability. Notably, content leakage appears to be closely related to layout disruptions. This can be validated from the experimental results of StyleCrafter [19] and IP-Adapter [38]. Although “A red

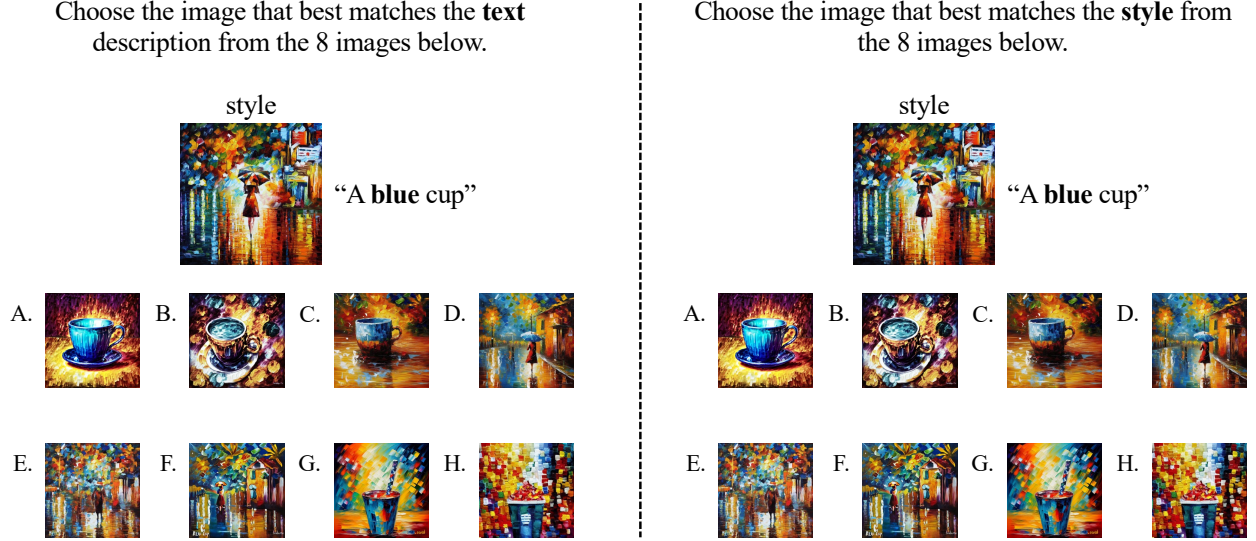


Figure 16. The questionnaire format for the user study. Each option represents the generation result of a method under a given style and prompt.

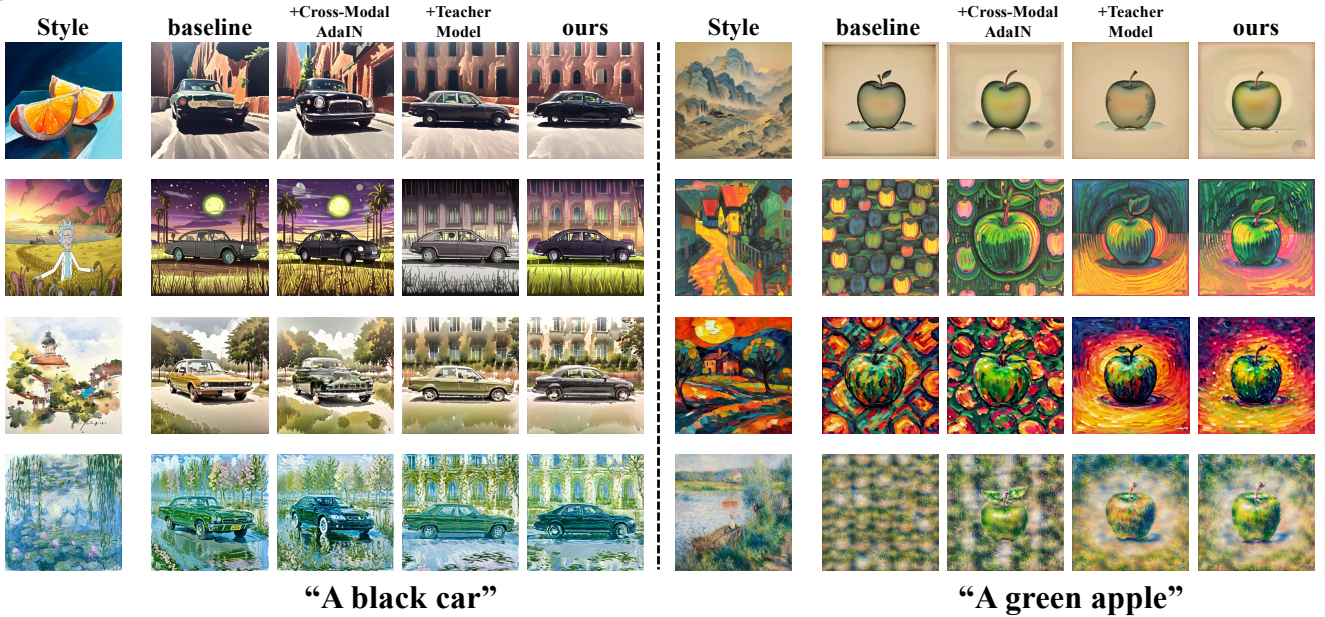


Figure 17. Qualitative results of the ablation study. cross-modal AdaIN enhances text alignment while preserving style similarity, addressing style overfitting issues. Incorporating the Teacher Model improves layout stability and resolves artifacts, ensuring consistent layout arrangements across different styles, as demonstrated in the “A green apple” example.

apple” is reflected in the final generated output, the image contains too many unrelated elements from the style reference, making it appear overly cluttered.

E. More results from our study

In Fig. 27, Fig. 28, and Fig. 15 we provide additional visualization results showcasing the effectiveness and versatility of our method. We have selected a variety of style categories and different color schemes to highlight the align-

ment effects for text descriptions. Moreover, we achieve excellent layout stability even when using the same prompt.

F. Integration with Other Methods

CSGO [37] has been recognized as one of the most effective and state-of-the-art methods for style transfer, which is why it was selected as the primary baseline in the main paper. To further evaluate the generalizability and robustness of our approach, we additionally explored its application and

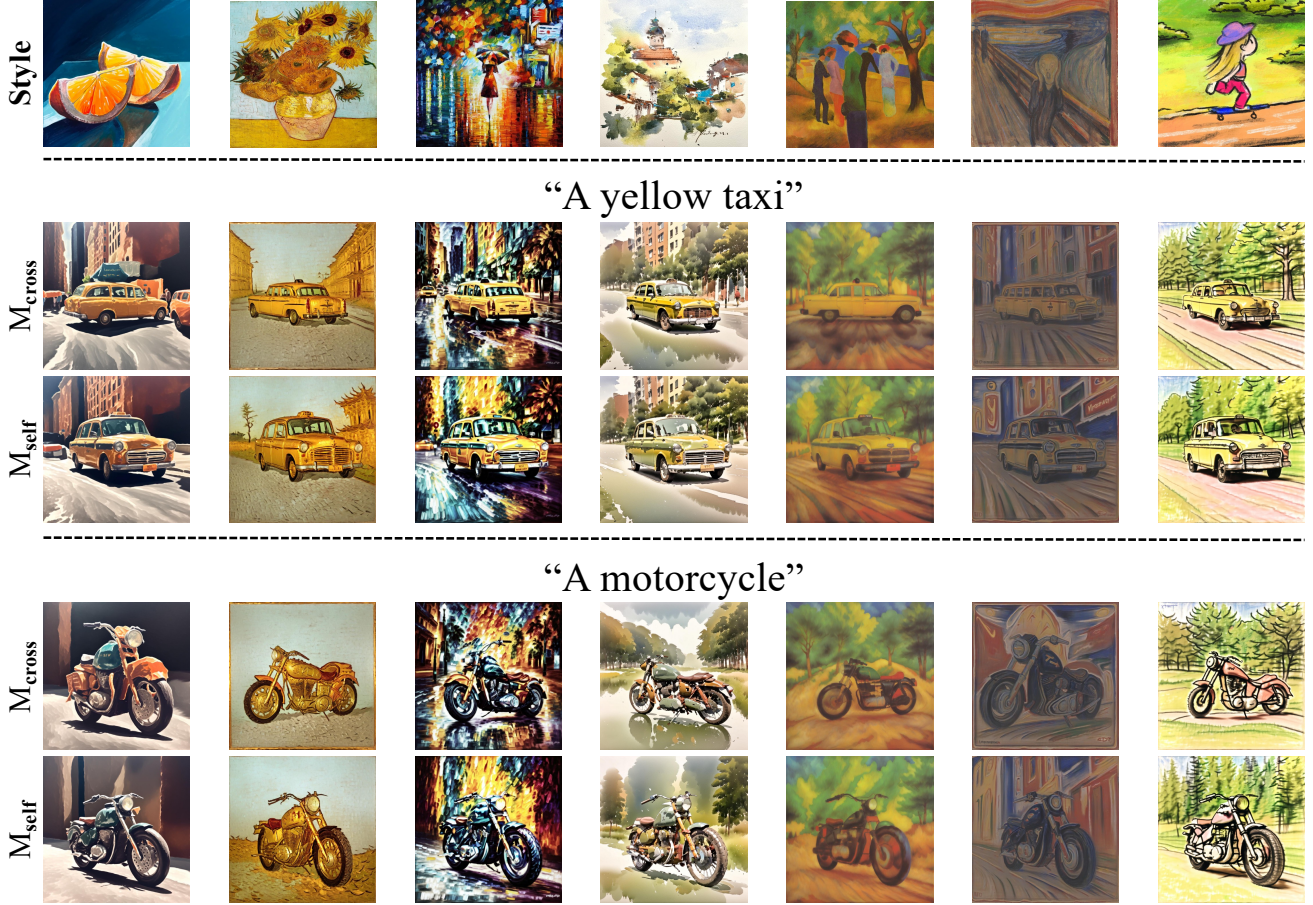


Figure 18. Implementation of the Teacher Model: Comparison of substituting the Self-Attention Map and Cross-Attention Map. The results demonstrate that replacing the Self-Attention Map achieves layout stability and consistency across different styles of images.

performance on other models.

F.1. Integration with InstantStyle [34]

Cross-Modal AdaIN. Since InstantStyle [34] is also an adapter-based architecture, it can similarly integrate cross-modal AdaIN to mitigate style overfitting. The results are shown in Fig. 21. Compared to Row 1, Row 2 accurately follows the text description, effectively avoiding errors in the generated output.

Teacher Model. InstantStyle [34] also encounters artifacts such as checkerboard patterns. Similar to the previous approach, we investigated the impact of the Teacher Model’s involvement at different timesteps on the results, as shown in Fig. 22. Upon observation, we reached a similar conclusion: if the Teacher Model participates for too many timesteps, it can lead to style loss.

F.2. Integration with StyleCrafter [19]

Teacher Model. A notable issue in StyleCrafter [19] is content leakage, where unrelated content elements from the style image appear in the generated results, ultimately affecting the final output. This phenomenon can lead to gen-

erated images that do not align with the descriptions in the prompt. To address this, we incorporated the Teacher Model into the method. As shown in Fig. 23, the inclusion of the Teacher Model significantly mitigates the problem of content leakage, resulting in outputs that maintain stability and consistency across different styles.

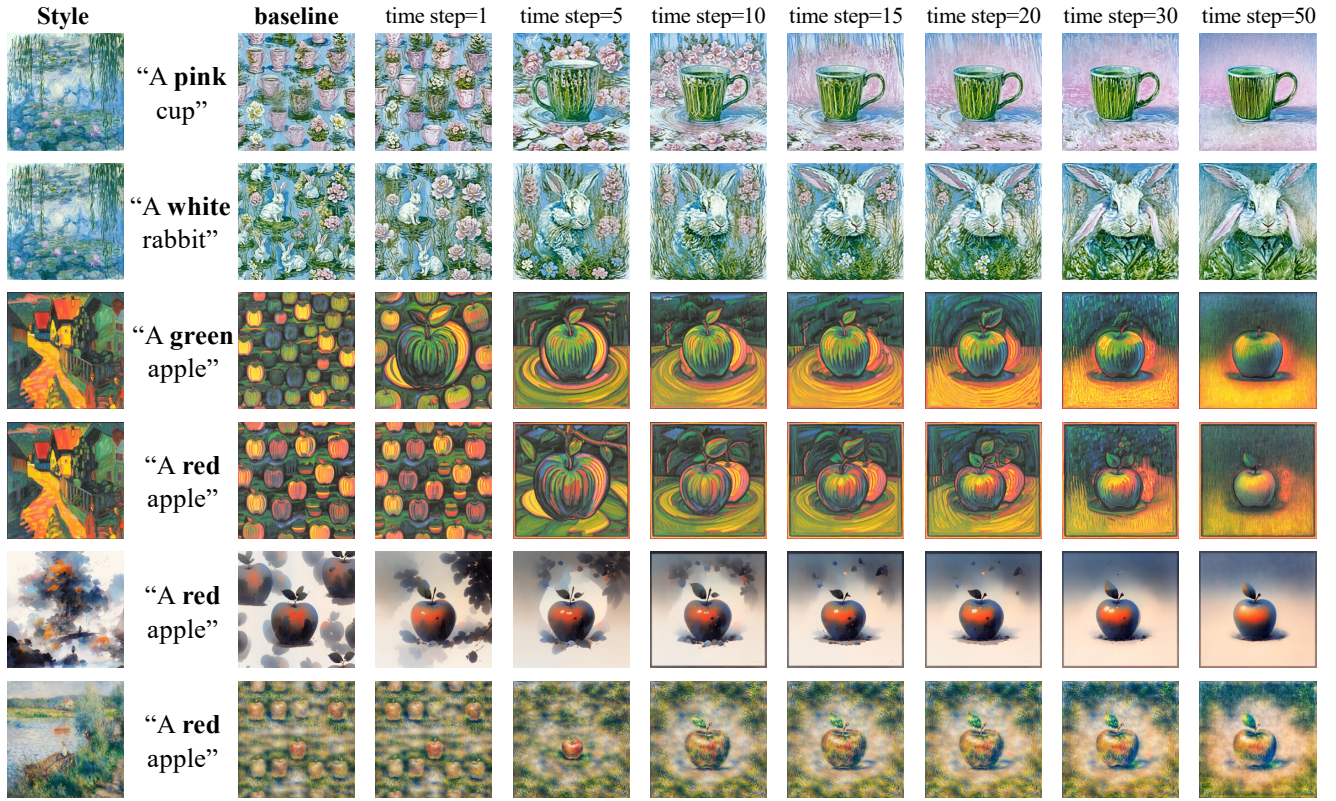


Figure 19. Impact of Teacher Model on Style Image Generation. The term “timestep” refers to the number of denoising steps during which the Teacher Model is involved. Notably, these experiments were conducted without incorporating cross-modal AdaIN to isolate and evaluate the specific impact of the Teacher Model on the generated results.



Figure 20. Compared to the image-based style transfer(I2I) provided by CSGO [37], We ensured the use of the same initial noise for both our method and the generation of the content image for I2I. It can be observed that the results obtained using the Teacher Model differ significantly from those of I2I, as I2I fails to preserve the color information of the original image.

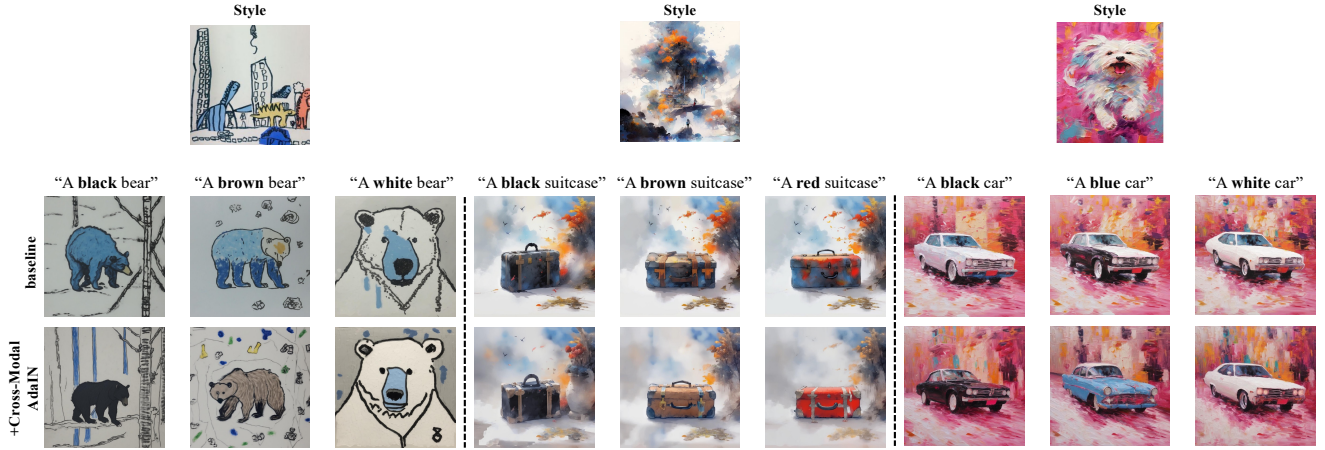


Figure 21. Qualitative results of using cross-modal AdaIN in InstantStyle [34]. The results demonstrate that cross-modal AdaIN effectively prevents style overfitting. The final generated results consistently align with the textual descriptions.

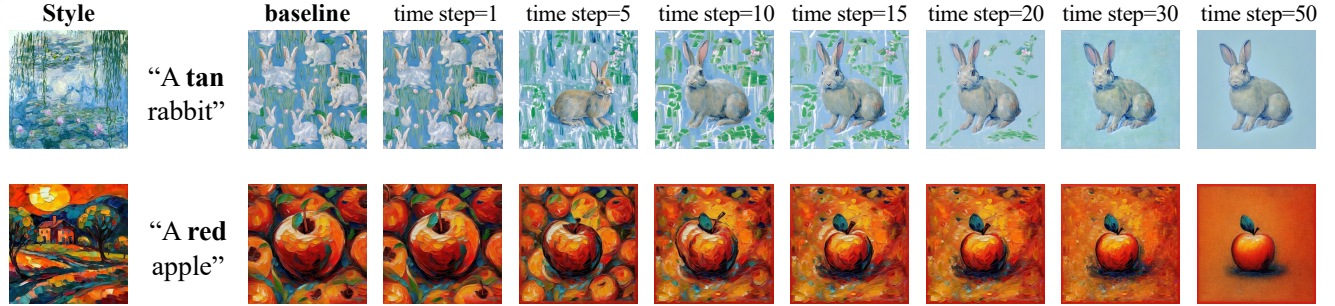


Figure 22. Impact of Teacher Model on InstantStyle [34] Image Generation. The term “timestep” refers to the number of denoising steps during which the Teacher Model is involved. Notably, these experiments were conducted without incorporating cross-modal AdaIN to isolate and evaluate the specific impact of the Teacher Model on the generated results. When the Teacher Model is applied to InstantStyle [34], it helps prevent the generation of artifacts, such as checkerboard patterns.

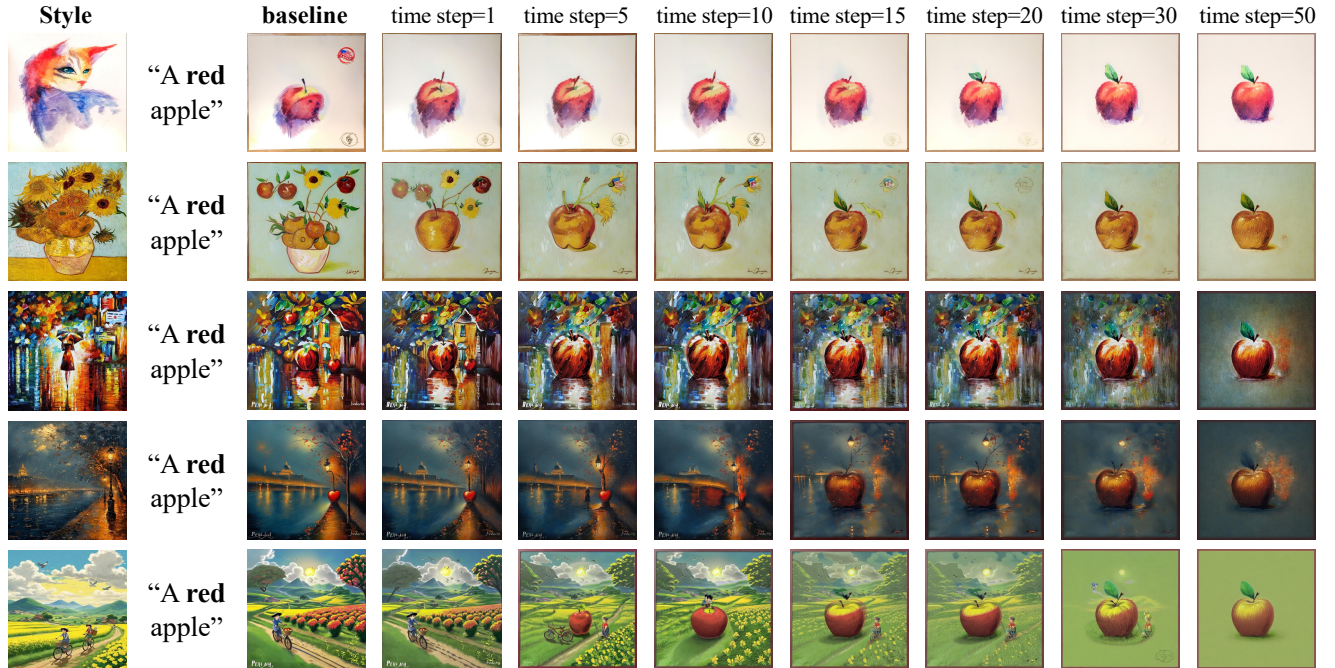


Figure 23. Impact of Teacher Model on StyleCrafter [19] Image Generation. The term “timestep” refers to the number of denoising steps during which the Teacher Model is involved. Notably, these experiments were conducted without incorporating cross-modal AdaIN to isolate and evaluate the specific impact of the Teacher Model on the generated results. In addition to ensuring layout stability, the Teacher Model also effectively reduces the occurrence of content leakage when applied to StyleCrafter [19].

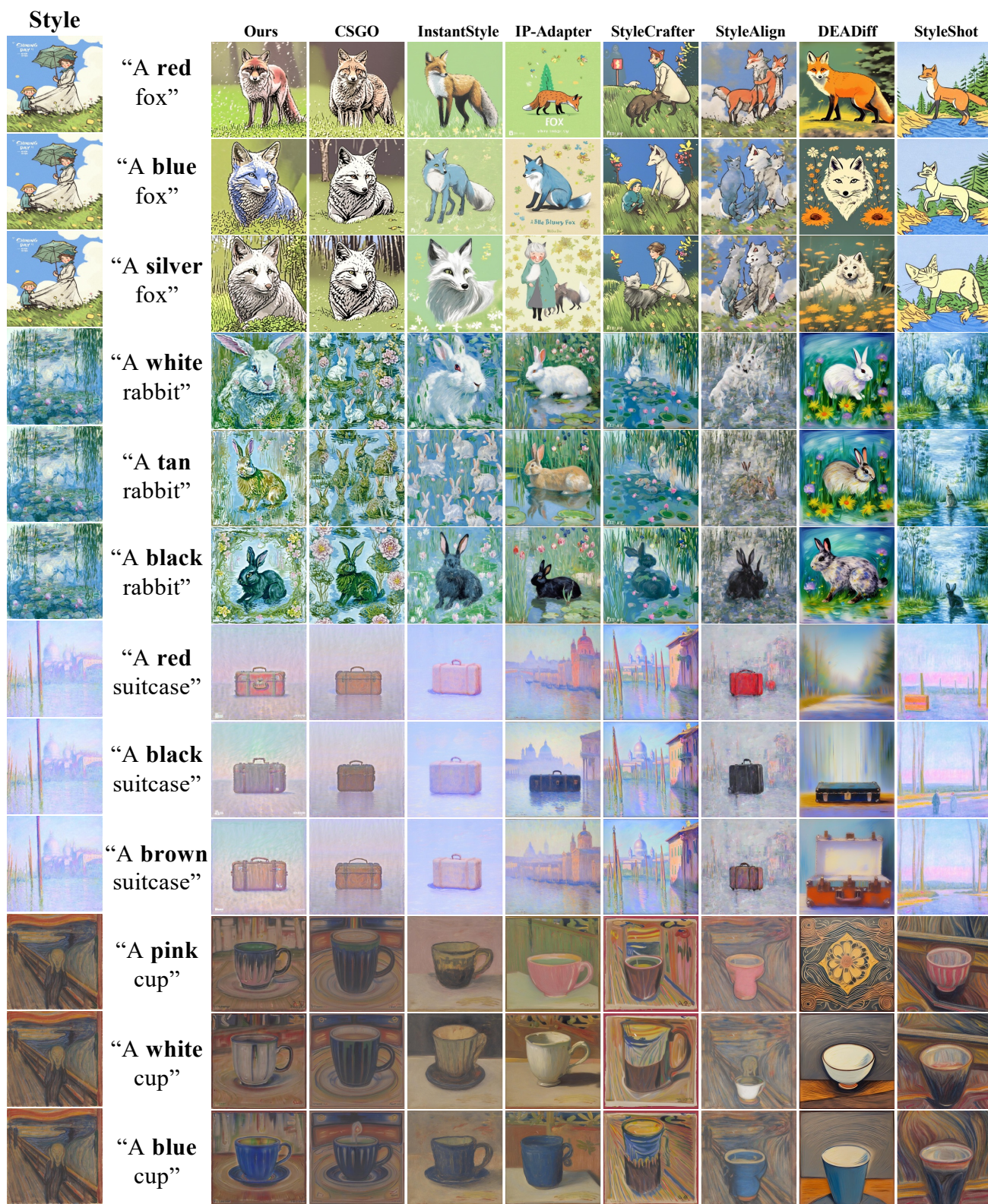


Figure 24. Qualitative comparison with state-of-the-art methods. Our approach effectively preserves image style while accurately adhering to text prompts for generation.



Figure 25. Qualitative comparison with state-of-the-art methods. Our approach effectively preserves image style while accurately adhering to text prompts for generation.

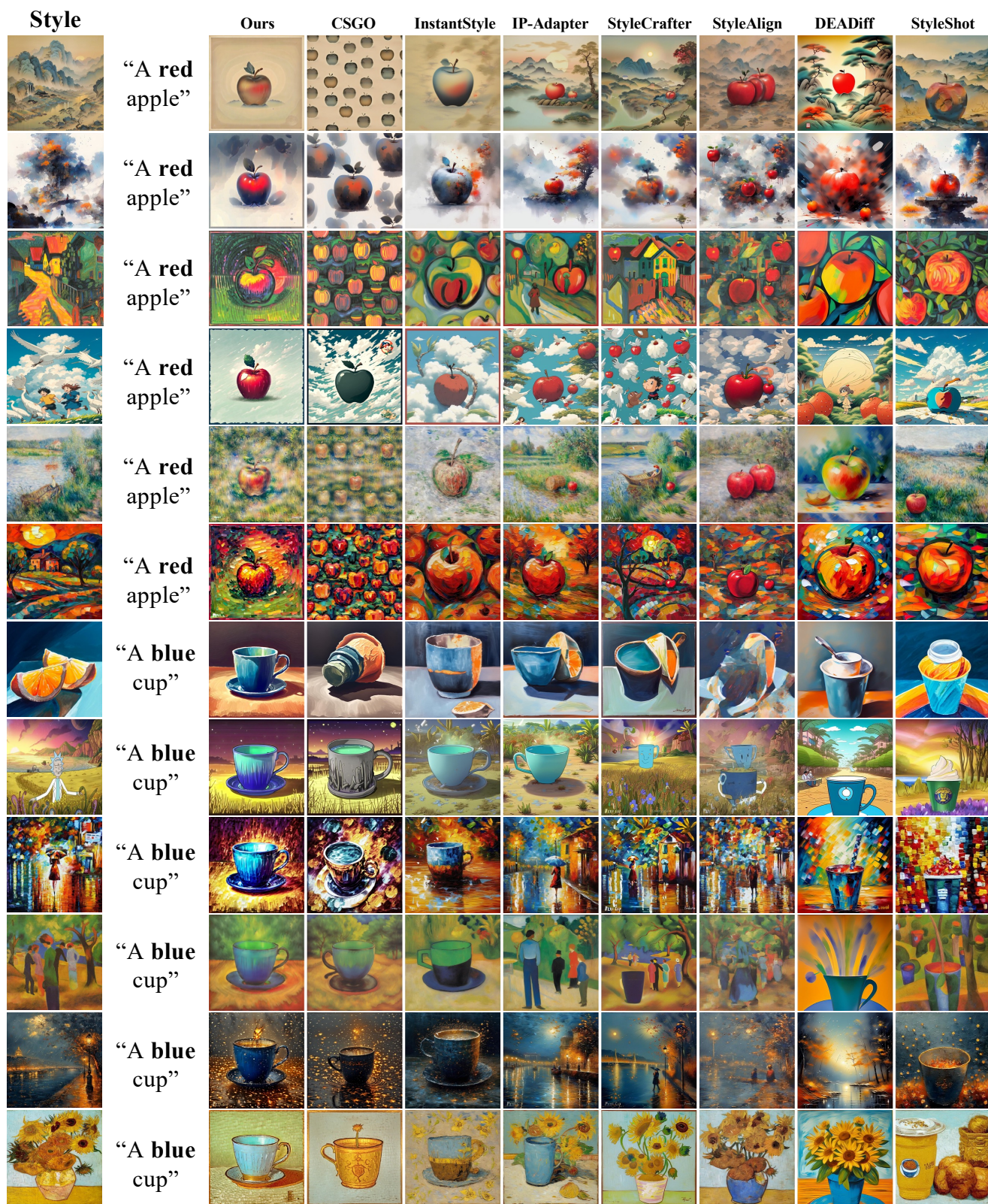
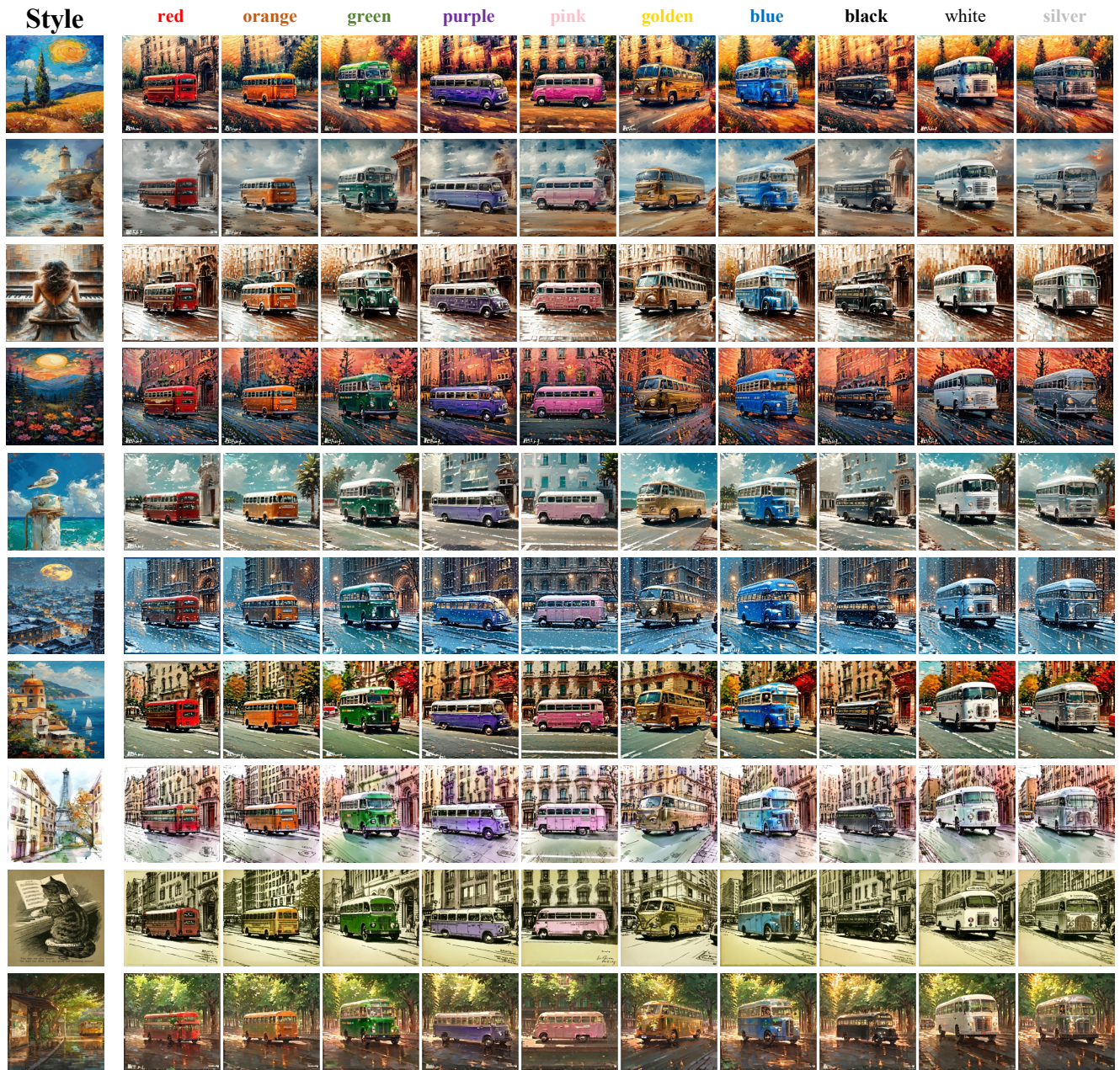


Figure 26. Qualitative comparison with state-of-the-art methods. Our approach effectively maintain layout consistency across different styles under the same prompt.



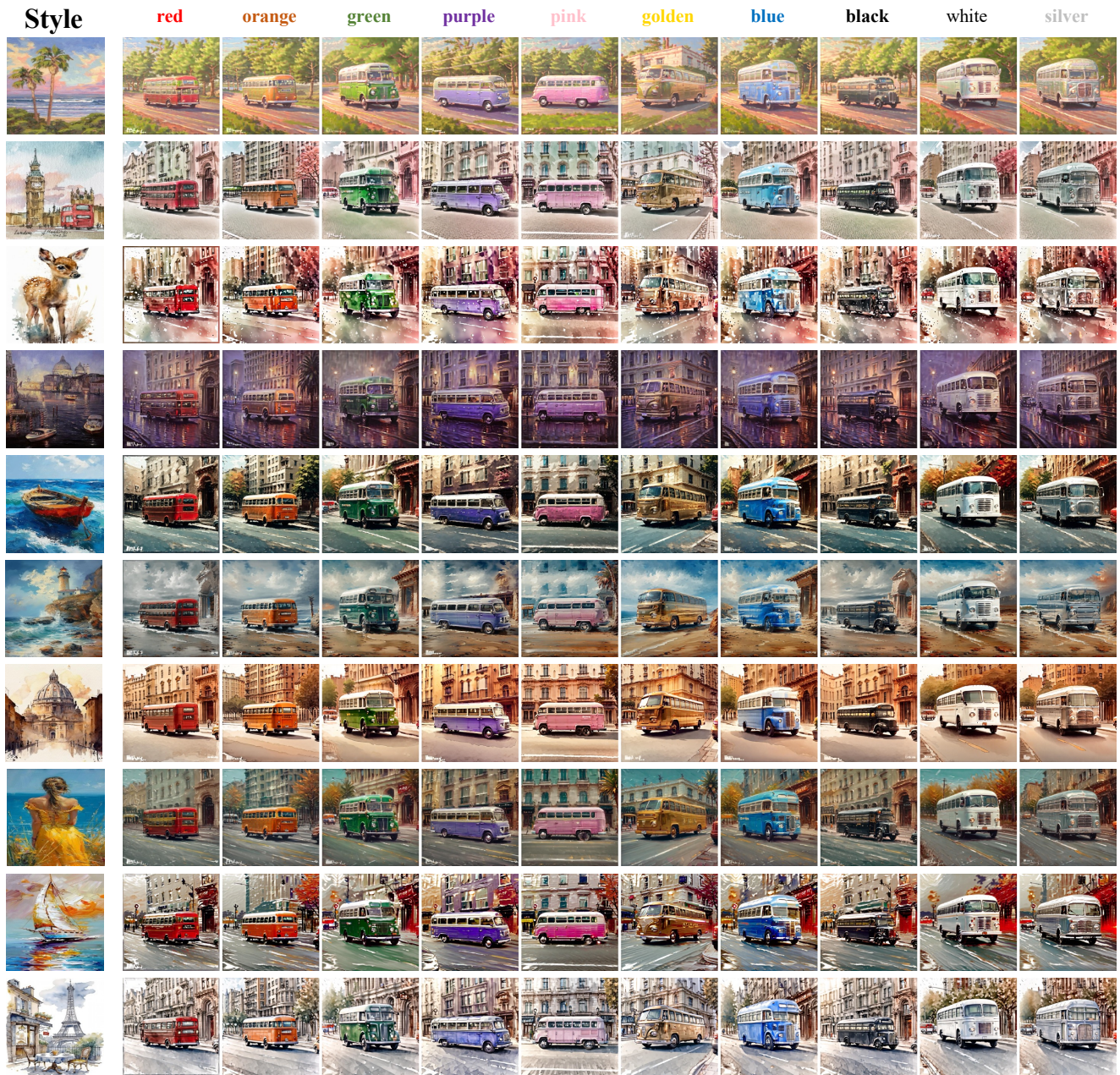


Figure 28. More results of our text-driven style transfer model. Given a style reference image, our method effectively reduces style overfitting, generating images that faithfully align with the text prompt while maintaining consistent layout structure across varying styles. Illustration of the prompt format used: “A [color] bus”.