AIComposer: Any Style and Content Image Composition via Feature Integration

Supplementary Material

1. Ablations and Additional analysis

1.1. Qualitative analysis of ablations

Figure 3 presents qualitative comparisons of our ablation study. The removal of inversion (-Inversion) exhibits noticeable degradation in foreground subject content features and introduces incoherence between the foreground and background. Similarly, the exclusion of ImageCLIP (-ImageCLIP) similarly demonstrates a substantial impact on the characteristics of the foreground subject. Additionally, -InitBlend significantly compromises the style transfer quality of foreground subjects.

1.2. Visualizing the features of MLP

In this work, we integrate the features of foreground content and background style with a simple MLP network. Therefore, the MLP network implicitly serves for style transfer. Although we can not visualize the pre-stylized image directly, the integrated features by MLP can be visualized with an IP-Adapter [15]. Figure 1 presents two typical examples of the composited results and the integrated features. We anticipate that the integrated feature captures the content of the foreground while pre-stylizing it referring to the background. In most cases, such as the top example

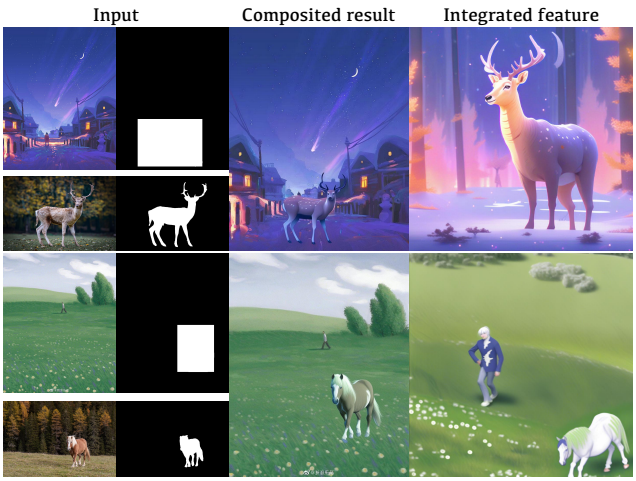

Figure 1. Two examples of the composited results and their corresponding integrated features by MLP (visualizing with IP-Adapter [15]). The top example is well pre-stylized, whereas the bottom example shows content leakage of a man in the integrated feature.

in Figure 1, the MLP network effectively performs implicit pre-stylizing. In other cases, such as the bottom example in Figure 1, even if the pre-stylizing features are not perfectly blended, our method still achieves robust results. This demonstrates the effectiveness of our feature blending strategy.

1.3. Effects of residual learning and number of triplet data

In this work, we employ $30,000$ triplets for training MLP in a **residual** manner, ensuring robust learning for various contents and styles. To learn the effect of the number of training data and the residual training strategy, we conduct additional experiments, including 300 and $3,000$ training data. We also evaluated the effect of the residual learning strategy by replacing it with direct learning for blended features. Figure 2 shows some examples of these settings. We visualize the integrated features by an IP-Adapter [15]. It is evident that the residual learning strategy contributes to robustness in the integration of the features of foreground content and background style. Figure 2 also indicates that the strong generalization performance of the feature blending strategy does not depend on a large number of training triplets. We attribute this success to our meticulous design of residual learning, which takes full consideration of the inherent additive properties of CLIP [9]. During inference, we add the CLIP features of the foreground and background images, and then subtract a learned residual. Even if the residual is not well learned (tending towards random features), our method can still effectively capture most of the content and style features from the input foreground and background images. This enhances the robustness of our method.

To validate the impact of different stylization methods on MLP performance, we additionally generated a subset of training data using StyleID [4], with corresponding quantitative results presented in Table 1. The results indicate that different stylization models have no significant impact on our method's performance, while data filtering and increased dataset size consistently improve results. This demonstrates the robustness of our approach across stylistic variations.

Table 1. Ablation study on the various data sources and stylization models. The best results are in bold, and the second best are underlined.

Data Num	Source	Filtered	LPIPS$_\downarrow$	CSD$_\uparrow$	CLIP-T$_\uparrow$
30K	CSGO [13]	Yes	**0.4195**	**0.5284**	**28.75**
300	CSGO [13]	Yes	0.4322	<u>0.5227</u>	28.05
300	CSGO [13]	No	0.4354	0.5159	28.01
30K	StyleID [4]	No	0.4340	0.5154	28.07
300	StyleID [4]	Yes	<u>0.4291</u>	0.5152	<u>28.21</u>
300	StyleID [4]	No	0.4327	0.5133	28.15

Figure 2. Examples of our composited results (the 2_{nd} column), and blended feature visualization (the last 6 columns) using different number of training data, with and without residual training strategy.

2. Results for same-domain composition

Despite its powerful capabilities in handling cross-domain image composition, our method can also be applied to same-domain cases. Table 2 summarizes the quantitative results on the same-domain benchmark of TF-ICON [6], as mentioned in the body text of this paper. We include the common CLIP similarity metrics for text prompts and foreground images, as well as LPIPS scores [17] for both foreground and background images. Our methods achieve significantly better LPIPS scores for the composed foreground, indicating superior preservation of foreground content over existing works. Some other metrics for our method are also competitive with the existing works. This demonstrates that our approach is robust in handling same-domain image compositions as well as the cross-domain cases detailed in the main text of this paper.

Table 2. Quantitative comparisons in the remaining 237 samples for the same-domain (photorealism) TF-ICON benchmark. The best results are in bold, and the second best are underlined.

Method	$\text{CLIP}_{(\text{Text})}\uparrow$	$\text{CLIP}_{(\text{Image})}\uparrow$	$\text{LPIPS}_{(\text{BG})}\downarrow$	$\text{LPIPS}_{(\text{FG})}\downarrow$
Blended [1]	25.19	73.25	0.11	0.77
Pa. by Ex. [14]	25.92	80.26	0.13	0.73
AnyDoor [3]	**31.24**	**87.87**	<u>0.09</u>	0.59
TF-ICON [6]	28.11	82.86	0.10	0.60
TALE [7]	31.03	<u>85.12</u>	0.10	0.51
PrimeCom. [11]	30.26	84.71	**0.08**	<u>0.48</u>
Ours	<u>31.08</u>	84.26	**0.08**	**0.36**

3. More results for cross-domain composition

We include 3 additional metrics to assess the results of different methods in our extended benchmark.

- **FID** [5] and **ArtFID** [12] measure the content and style similarity between the resulting image and the hybrid image (by directly pasting the foreground image onto the background image) within the background masked region M_{bg}.
- **CLIP-I** [9] similarity measures the alignment score between the background masked region of the resulting image and the foreground image.

The FID and ArtFID metrics are also effective in measuring the composited results in both content preservation and style consistency. CLIP-I acts similarly to CLIP-T, which is commonly used in same-domain image composition. The quantitative results for different methods are included in Table 3. We can see that our method suppresses the existing ones except for the CLIP-I score. We argue that the CLIP-I metric focuses more on content preservation and does not account for style transformation.

We also implement our method with an early version of stable diffusion (*ours-SD1.5*) [10] for more comprehensive evaluations. Although an advanced SDXL model [8] leads to some gains in quantitative performance, combining our method with an early version of SD also outperforms the existing works by a significant margin. This demonstrates that the enhanced performance of our method is primarily due to MLP feature blending and single-branch diffusion strategies.

More qualitative results are in Figure 7, Figure 8, and Figure 9. With these versatile examples, we hope that readers can gain a more intuitive understanding of our methods and our proposed benchmark dataset for cross-domain image composition.

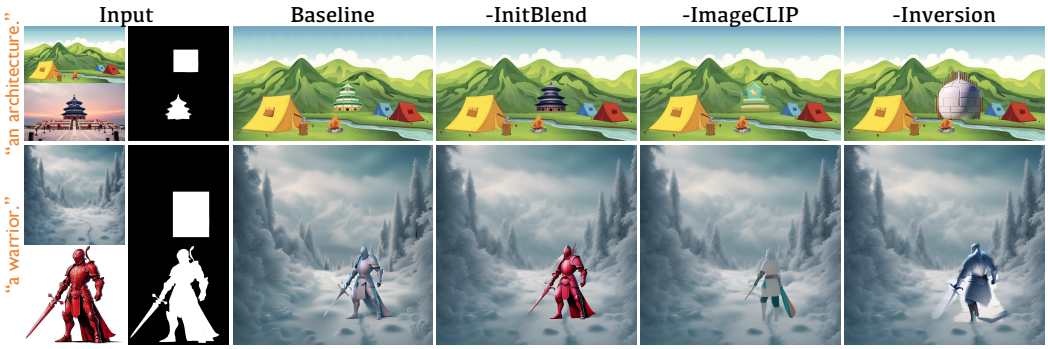

Figure 3. The qualitative results of ablation study.

Table 3. Quantitative comparisons in the extended benchmark. The best results are in bold, and the second best are underlined. We also include an early version of stable diffusion (Ours-SD1.5) and the null-prompt version for our method (Ours-NP).

Method	BaseModel	LPIPS$_\downarrow$	FID$_\downarrow$	ArtFID$_\downarrow$	CSD$_\uparrow$	PSNR$_\uparrow$	CLIP-T$_\uparrow$	CLIP-I$_\uparrow$
Blended [1]	SDXL	0.6743	13.77	24.74	<u>0.4473</u>	<u>18.92</u>	25.87	66.33
Paint by Example [14]	SD1.4	0.6684	12.29	22.17	0.3175	15.14	28.73	<u>78.94</u>
AnyDoor [3]	SD2.1	<u>0.6036</u>	<u>9.587</u>	<u>16.95</u>	0.2963	18.24	<u>29.02</u>	**84.40**
TF-ICON [6]	SD2.1	0.6707	12.82	23.05	0.4013	13.62	28.23	70.98
Ours	SDXL	**0.4195**	**9.538**	**14.97**	**0.5283**	**19.48**	**29.29**	77.59
Ours-NP	SDXL	0.4221	9.682	15.20	0.5294	19.49	28.75	75.93
Ours-SD1.5	SD1.5	0.5093	11.15	18.31	0.4886	16.63	28.79	74.40

4. Extensions

4.1. Combination with ControlNet

The proposed method can be readily combined with ControlNet [16], since it involves no modification of the main diffusion process. Figure 4 presents two examples with and without ControlNet. We use the canny edges [2] of the foreground images for additional control. In some cases, such as the top example in Figure 4, this conditional strategy enhances the details. However, in other cases, such as the bottom example, rigid control with canny edges may undermine the style coherence of the foreground. Additionally, integrating ControlNet introduces extra computational costs.

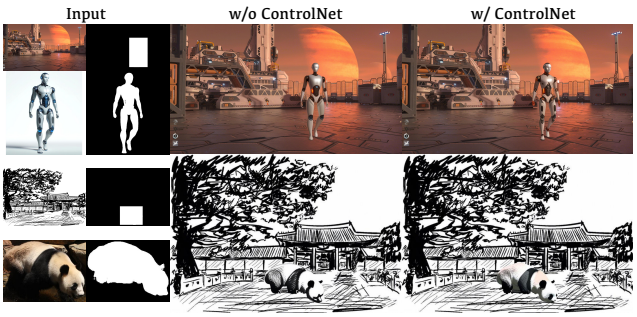

Figure 4. Two examples of our method with and without ControlNet. The top example demonstrates a positive effect on enhancing the details of the robot, while the bottom example shows a negative effect by hindering the stylization.

4.2. Exchanging foreground and background

Another distinguishing feature of our method, compared to existing methods, is its **symmetric** treatment of foreground and background images. This symmetry allows for straightforward generalization to various tasks, such as image stitching where the roles of input images can be easily interchanged. We can also exchange the role of the foreground and background, *i.e*. blending the background into the foreground. Figure 5 (right column) shows an example. We can see that the background is well stylized while the foreground remains exactly unchanged. This task is somewhat different from the image composition task (left) and is useful in practical applications.

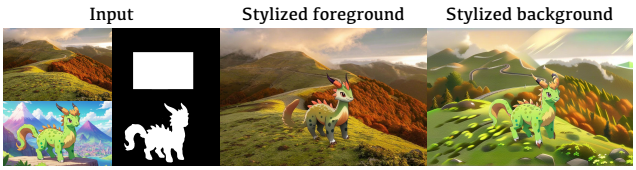

Figure 5. An example of exchanging the roles of the foreground and the background images. Our method allows easy stylization of the foreground with reference to the background, and vice versa.

5. More details

MLP network architecture. Figure 6 shows a simple implementation of the MLP network architecture in PyTorch.

```
class MLPCLIPFeatureNetwork(nn.Module):
    def __init__(self, input_dim=2*2048*8, hidden_dim=512,
    output_dim=2048*8):
        super(MLPCLIPFeatureNetwork, self).__init__()
        self.mlp = nn.Sequential(
            nn.Linear(input_dim, hidden_dim),
            nn.ReLU(),
            nn.Linear(hidden_dim, hidden_dim),
            nn.ReLU(),
            nn.Linear(hidden_dim, output_dim),
        )
    def forward(self, input_features):
        return self.mlp(input_features)
```

Figure 6. Codes for the 3-layer MLP network.

Principles of Filtering Training Data for MLP. We generate content-style stylized triplets by leveraging a state-of-the-art neural-style transfer method [13], resulting in a preliminary collection of $65,429$ triplets. The quality and erroneous samples are evaluated on three criteria.

(a) **Content inconsistency**: Removal of triplets where the primary subject in the stylized output deviates significantly from the content image.

(b) **Style discrepancy**: Exclusion of the stylized results exhibiting mismatches in critical style attributes (e.g., color palette, texture patterns, and brushstroke characteristics) compared to the style image.

(c) **Substandard visual quality**: Filter artifact-containing outputs with noticeable distortions, blurring, or degradation in perceptual fidelity.

Finally, we get $37,445$ rigorously filtered style transfer instances.

Questions in user study. We randomly shuffle the results from different methods to ensure anonymity in our user study. Given the foreground and background images, as well as their corresponding masks, we ask the users to select the best results referring to the following questions.

(a) **Content preservation**: The synthetic image should retain the characteristics of the foreground subject (such as its identity, shape, and outline) well, without considering color changes.

(b) **Style consistency**: The style of the foreground subject in the composite image should be consistent with the style of background image, taking into account elements such as tone, texture, and line patterns.

(c) **Seamless blending**: The composite image needs to seamlessly blend the foreground and background, making it difficult to detect any signs of stitching. There should be no visible edges, artifacts, irregular color patches, or obvious mapping traces.

Computation of CLIP-T. In the samples from the TF-ICON [6] dataset, where the text prompt describes global image information, we compare the alignment score between the entire resulting image and the given text prompt. However, in most examples from our new benchmark, we compare the alignment score between the locally masked region of the resulting image and the given text prompt. This approach considers that the text prompt, such as 'a dog', describes local information.

References

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Trans. Graph.*, 42(4):149:1–149:11, 2023. 2, 3

[2] John F. Canny. A computational approach to edge detection. *IEEE TPAMI*, 8(6):679–698, 1986. 3

[3] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, pages 6593–6602, 2024. 2, 3

[4] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *CVPR*, pages 8795–8805, 2024. 1

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017. 2

[6] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. TF-ICON: diffusion-based training-free cross-domain image composition. In *ICCV*, pages 2294–2305, 2023. 2, 3, 4

[7] Kien T. Pham, Jingye Chen, and Qifeng Chen. TALE: training-free cross-domain image composition via adaptive latent manipulation and energy-guided optimization. In *ACM MM*, pages 3160–3169, 2024. 2

[8] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 2

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 2

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685, 2022. 2

[11] Yibin Wang, Weizhong Zhang, Jianwei Zheng, and Cheng Jin. Primecomposer: Faster progressively combined diffusion for image composition with attention steering. In *ACM MM*, pages 10824–10832, 2024. 2

[12] Matthias Wright and Björn Ommer. Artfid: Quantitative evaluation of neural style transfer. In *DAGM German Conference on Pattern Recognition*, pages 560–576, 2022. 2

[13] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li.

Figure 7. Qualitative comparison of our method with prior SOTA works. We also include an early version of stable diffusion (Ours-SD1.5) and the null-prompt version for our method (Ours-NP).

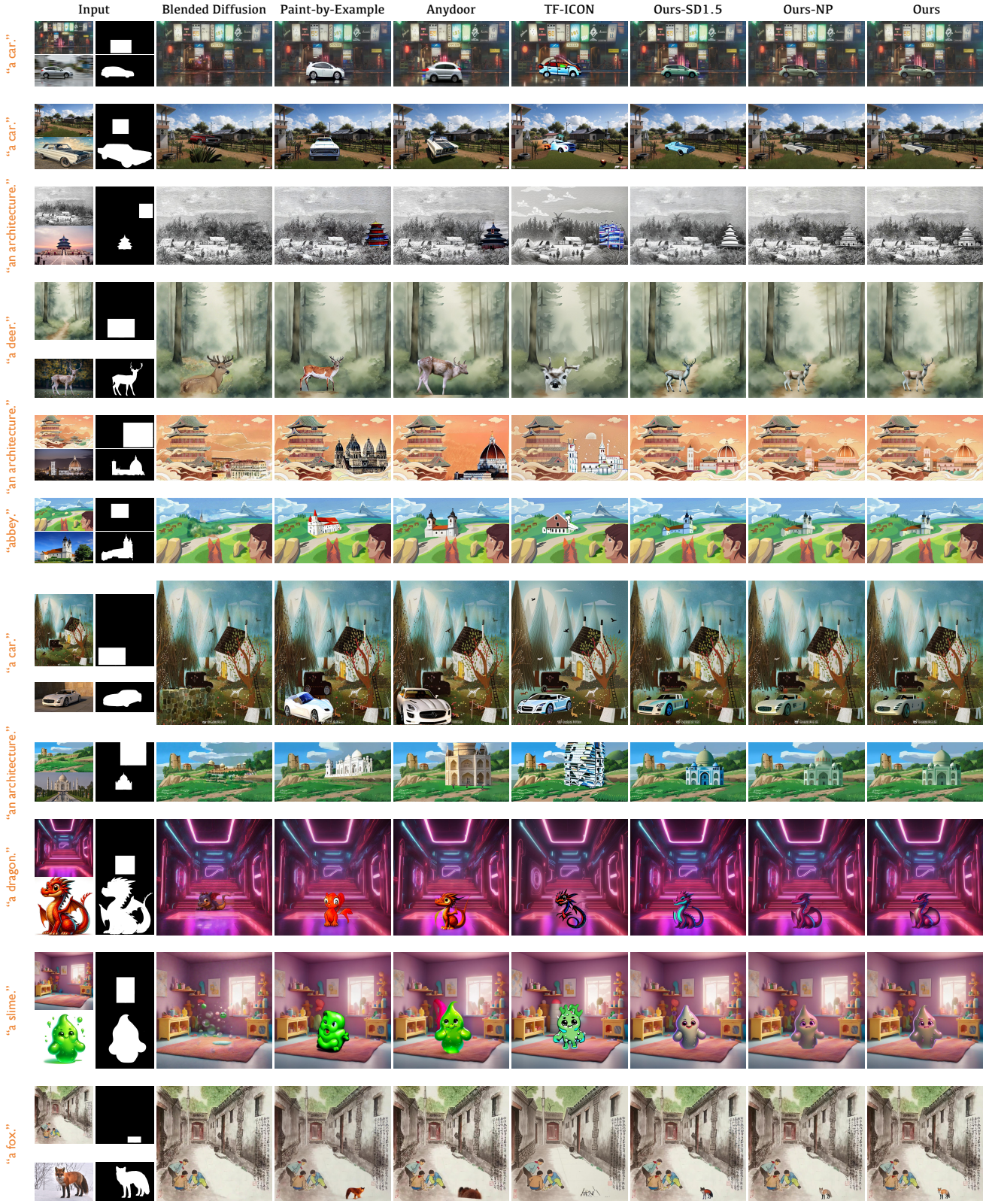

Figure 8. Qualitative comparison of our method with prior SOTA works. We also include an early version of stable diffusion (Ours-SD1.5) and the null-prompt version for our method (Ours-NP).

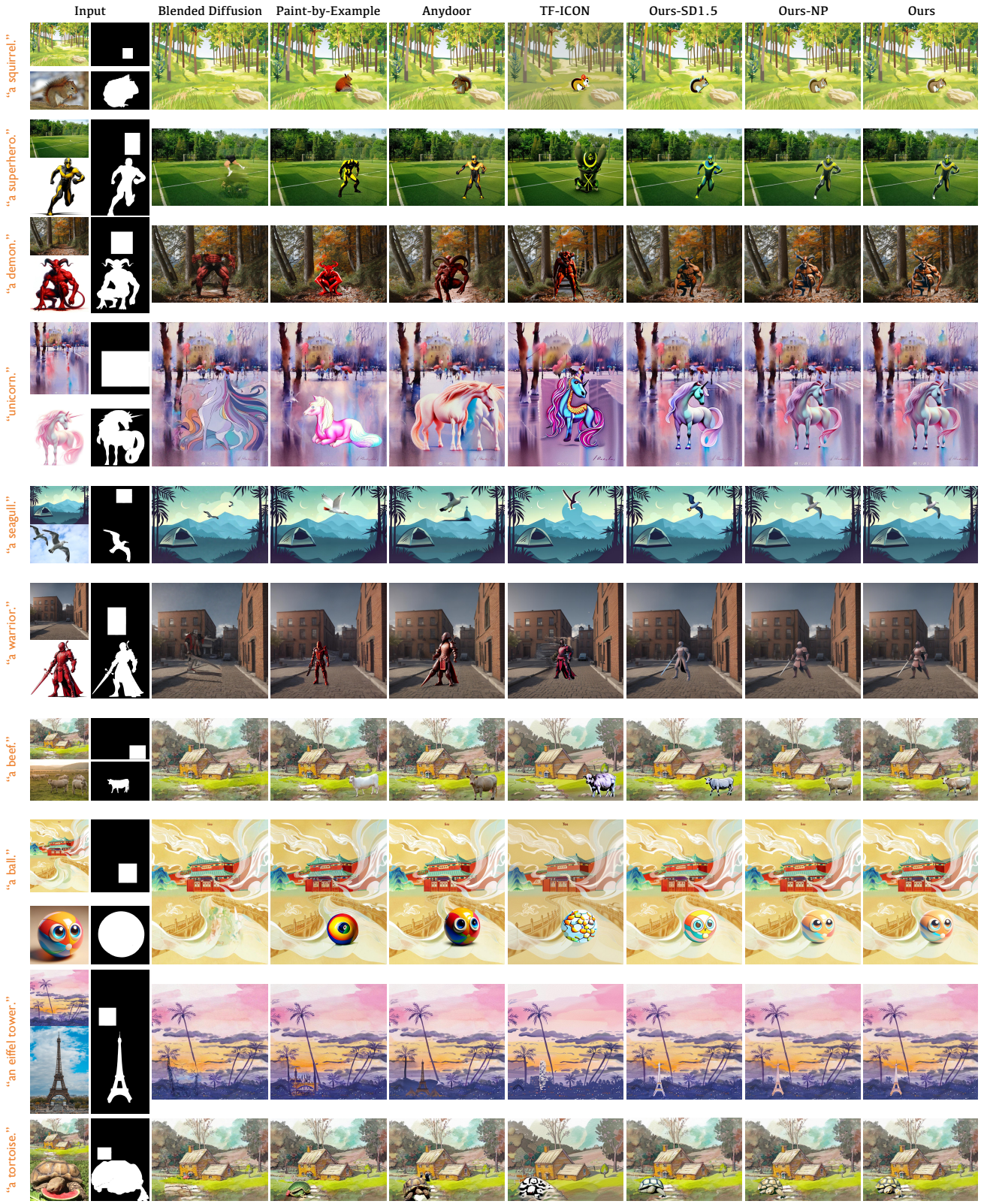

Figure 9. Qualitative comparison of our method with prior SOTA works. We also include an early version of stable diffusion (Ours-SD1.5) and the null-prompt version for our method (Ours-NP).

CSGO: content-style composition in text-to-image generation. *CoRR*, abs/2408.16766, 2024. 1, 4

[14] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*, pages 18381–18391, 2023. 2, 3

[15] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *CoRR*, abs/2308.06721, 2023. 1

[16] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3813–3824, 2023. 3

[17] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 2