UNIC-Adapter: Unified Image-instruction Adapter with Multi-modal Transformer for Image Generation

Supplementary Material

1. Additional implementation details

In this section, we provide additional implementation details of our UNIC-Adapter.

1.1. Model Architecture

As described in the main paper, our UNIC-Adapter shares the same architecture as the SD3 medium model [2] and is initialized using the parameters of the SD3 medium. Specifically, our UNIC-Adapter consists of 24 MM-DiT blocks, with each block containing two AdaLayerNormZero layers, one Attention layer, and two Feed-Forward layers. To reduce the number of trainable parameters, we freeze the parameters of the Feed-Forward layers and only train the remaining layers.

1.2. Training Details

The dataset mixing ratios are set as follows: pixel-level spatial control: 0.4, subject-driven image generation: 0.5, and style-image-based image generation: 0.1. For subject-driven image generation, the background of the subject images is set to white. The input images are first resized so that the shorter side is 512 pixels, and then they are randomly cropped to a resolution of 512×512. To enable classifier-free guidance, for pixel-level spatial control, we use a probability of 0.15 to drop the text prompt. For the other two tasks, we use the following probabilities: 0.05 to drop the text prompt, 0.05 to drop both the task instruction and the conditional image simultaneously, and 0.05 to drop the text prompt, task instruction, and conditional image simultaneously. Our UNIC-Adapter is trained using the same loss function as SD3 medium [2].

1.3. Inference Details

During inference, we use the same sampling schedule as SD3, with the sampling step set to 28. We employ classifier-free guidance based on three conditions: the text prompt c_{txt}, the task instruction c_{ist}, and the conditional image c_{con}. The classifier-free guidance is performed as follows:

$$\begin{aligned} &e_\theta(z_t, c_{txt}, c_{ist}, c_{con}) = e_\theta(z_t, \varnothing, \varnothing, \varnothing) \\ &+ s_c \cdot (e_\theta(z_t, \varnothing, c_{ist}, c_{con}) - e_\theta(z_t, \varnothing, \varnothing, \varnothing)) \\ &+ s_t \cdot (e_\theta(z_t, c_{txt}, c_{ist}, c_{con}) - e_\theta(z_t, \varnothing, c_{ist}, c_{con})) \end{aligned} \tag{1}$$

where e_θ denotes the model, z_t denotes the image latents, $\varnothing$ denotes the fixed null value, s_c is the scale for image-instruction guidance, and s_t is the scale for text prompt guidance. For pixel-level spatial control, s_c and s_t are set to 1.3 and 3.0, respectively. For subject-driven image generation, s_c and s_t are set to 1.2 and 7.5, respectively. For style-image-based generation, s_c and s_t are set to 3.0 and 6.0, respectively.

2. Additional Experimental Results

In this section, we present additional experimental results, including both quantitative ablation studies and qualitative evaluations.

2.1. More Quantitative Results

Importance of Cross-modal Interaction Our UNIC-Adapter leverages the MM-DiT block, where task instruction features and conditional image features mutually attend to each other. To investigate the importance of this cross-modal interaction, we perform an experiment by modifying the attention process in Equation (4) of the main paper to align with the DiT block formulation [1]. The modified process is defined as follows:

$$\begin{aligned} &K_{\text{ist}} = L^k_{\text{ist}}(Z_{\text{ist}}), V_{\text{ist}} = L^v_{\text{ist}}(Z_{\text{ist}}), \\ &Q_{\text{con}} = L^q_{\text{con}}(Z_{\text{con}}), K_{\text{con}} = L^k_{\text{con}}(Z_{\text{con}}), V_{\text{con}} = L^v_{\text{con}}(Z_{\text{con}}), \\ &Z'_{\text{con}} = \text{Attn}(Q_{\text{con}}, [K_{\text{ist}} \| K_{\text{con}}], [V_{\text{ist}} \| V_{\text{con}}]), \end{aligned} \tag{2}$$

where the task instruction features no longer attend to conditional image features and instead act solely as key and value features without being updated. As shown in Table 1, removing cross-modal interaction leads to a decline in performance across several metrics, emphasizing the advantage of such interaction between task instruction features and conditional image features.

Comparison with UniControl. We conducted further quantitative evaluations using CLIP-I and FID metrics across 5,000 images within the rest eight pixel-level tasks. As presented in Table 2, our UNIC-Adapter consistently outperforms UniControl [5], the only comparative method spanning these tasks.

Effect of the Data Ratios. Table 3 shows how multi-task training affects performance by reporting the score on three pixel-level tasks (Canny, HED, and Depth): With an equal number of training steps, models trained on more tasks perform slightly worse than those trained on 3 pixel-level tasks. And varying the data ratios between 12 pixel-level and subject tasks has minimal impact on performance.

2.2. Scalability Experiments

Our UNIC-Adapter demonstrates remarkable flexibility in accommodating various conditional images and instruc-

Method	Canny (F1 Score↑)	HED (SSIM↑)	Seg. (mIoU↑)	Depth (RMSE↓)	Subject-Driven DINO↑	CLIP-I↑	CLIP-T↑
UNIC-Adapter	37.95	0.8420	33.32	32.25	0.784	0.829	0.309
w/o cross-modal	37.71	0.8284	31.73	31.95	0.772	0.821	0.310

Table 1. Results of UNIC-Adapter with and without cross-modal interaction between task instruction features and conditional image features on pixel-level control tasks and subject-driven generation task.

Method	Bbox	Blur	Colorization	Sketch	Inpainting	Normal	Skeleton	Outpainting
UniControl	78.3/17.5	94.7/**9.3**	91.9/10.0	79.2/21.4	92.6/10.3	84.3/16.7	78.6/17.9	86.6/13.5
Ours	**79.0/17.0**	**96.0**/11.3	**93.8/8.1**	**86.5/15.5**	**94.9/8.2**	**85.7/14.1**	**79.5/17.0**	**89.2/9.7**

Table 2. CLIP-I(↑) / FID(↓) scores on eight pixel-level control tasks.

Task metrics	3 pixel-level tasks	12 pixel-level tasks	12 pixel-level and subject tasks (0.75:0.25)	(0.5:0.5)	(0.25:0.75)
Canny (F1 Score↑)	41.66	38.74	35.65	36.17	36.90
HED (SSIM↑)	85.93	85.76	82.46	82.83	81.93
Depth (RMSE↓)	30.63	32.36	31.07	31.50	32.29

Table 3. Results with different training tasks and data mixing ratios.

tions. To further assess its scalability, we conducted experiments involving three distinct types of data: multiple conditional images, image editing, and subject-driven generation using subject images with backgrounds.

Multiple Conditional Images. As outlined in Eq. 5 of the main paper, our UNIC-Adapter inherently supports multiple conditional images by aggregating the attention results from each conditional input. Figure 1 illustrates zero-shot inference results using multiple conditional images, achieved with a model trained on single-condition input. This demonstrates the model's potential for enhanced performance when trained on a large-scale dataset with multiple conditional images.

Image Editing. We trained our model on the OmniEdit dataset [7], enabling it to receive both the original image and an instruction prompt detailing the desired modifications. Figure 2 demonstrates the effective editing capabilities of our UNIC-Adapter. In the future, we could further integrate the image editing task along with all pixel-level and subject-specific tasks into our framework.

Subject-driven Generation. In the subject-driven image generation task, the generated subjects exhibit limited variations in pose compared to the subject images, since the subject image and target image originate from the same image source, and the backgrounds in the subject image are erased during training, as Kosmos-G [4] did. To showcase that our model can also support subject-driven generation using subject images with backgrounds, we fine-tuned it on the Subjects200K dataset [6], which includes subject images with various backgrounds. As shown in Figure 3, our model effectively handles data with backgrounds and has comparable capabilities to OmniGen [8], generating diverse subject images across varying contexts during testing.

2.3. More Qualitative Results

Figures 4, 5, 6, and 7 showcase additional visualization results of our UNIC-Adapter across various controllable generation tasks.

3. Limitations and Future Work

In the experiments, all training images are resized and cropped to a resolution of 512×512. As a result, our UNIC-Adapter is limited in generating images with higher resolution, like 1024×1024. This limitation can be addressed by using high-resolution training images and keeping the original aspect ratio, such as images with pixel areas equivalent to 1024×1024. Additionally, although our model can support image generation with multiple conditional images during inference, it is limited by the training data which only contain one image as the condition. In the future, we can construct large-scale datasets with multiple conditional images to improve the capabilities. Furthermore, integrating our UNIC-Adapter with state-of-the-art T2I models, such as FLUX1.0-dev [3] and Stable Diffusion 3.5 Large [2], might further enhance the controllability and performance of these models.

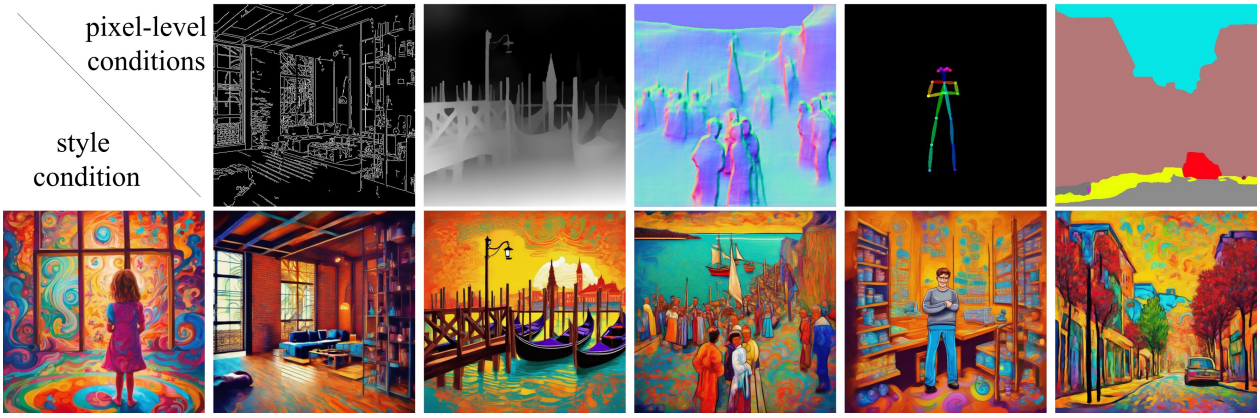
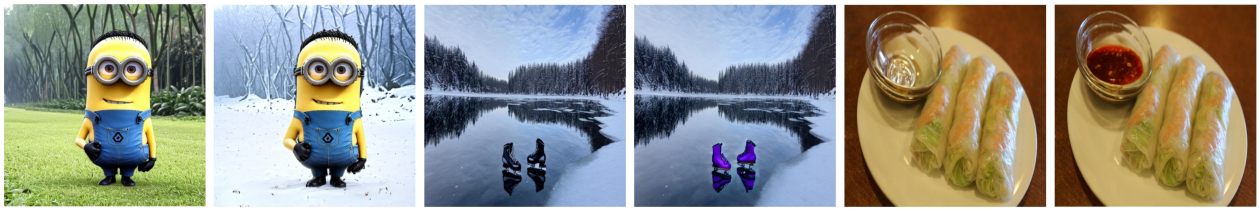


Figure 1. Zero-shot inference results of our UNIC-Adapter with multiple conditions.



change the setting to a snowy scene *turn the color of a set of ice skates to purple* *add chili sauce to the small glass bowl*

Figure 2. Image editing results of our UNIC-Adapter on OmniEdit test set.

Figure 3. Visualization results of OmniGen and our UNIC-Adapter on DreamBench. The first row is subject images with background, while the second row is generated images.

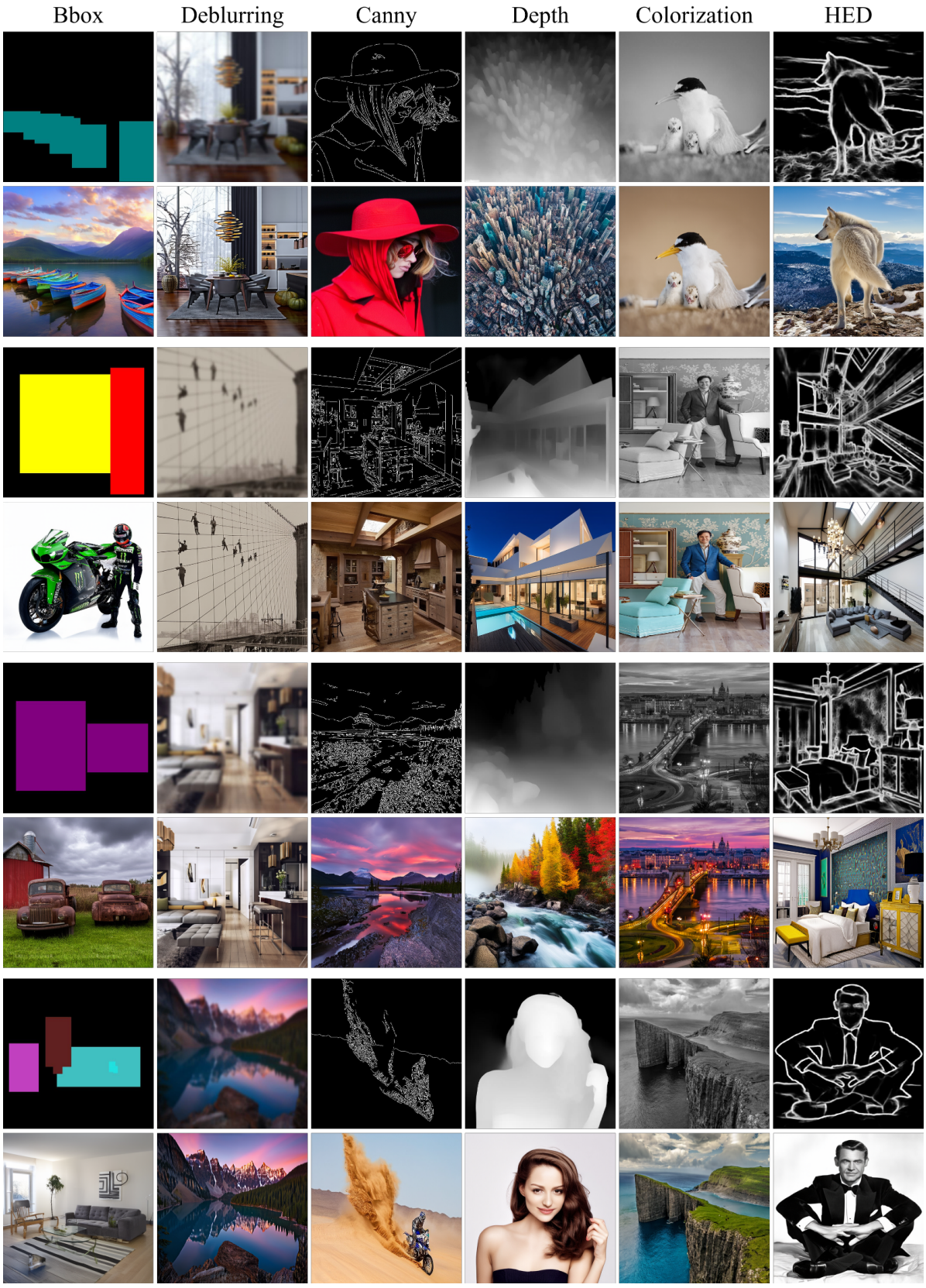


Figure 4. Visualization results of our UNIC-Adapter on six pixel-level control tasks from the MultiGen-20M dataset. The odd rows show different types of conditional images, while the even rows display the corresponding generated images.

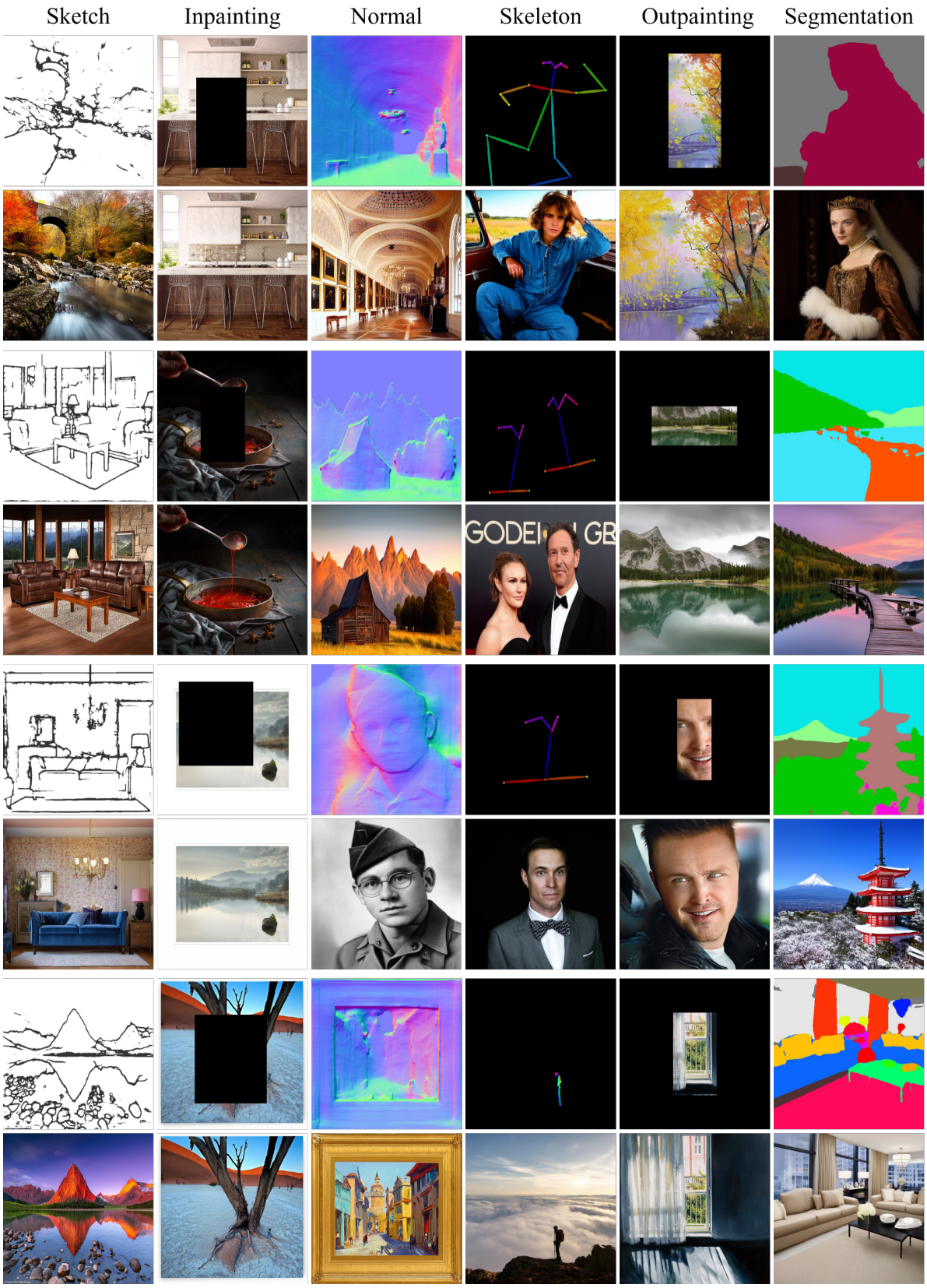


Figure 5. Visualization results of our UNIC-Adapter on six pixel-level control tasks from the MultiGen-20M dataset. The odd rows show different types of conditional images, while the even rows display the corresponding generated images.

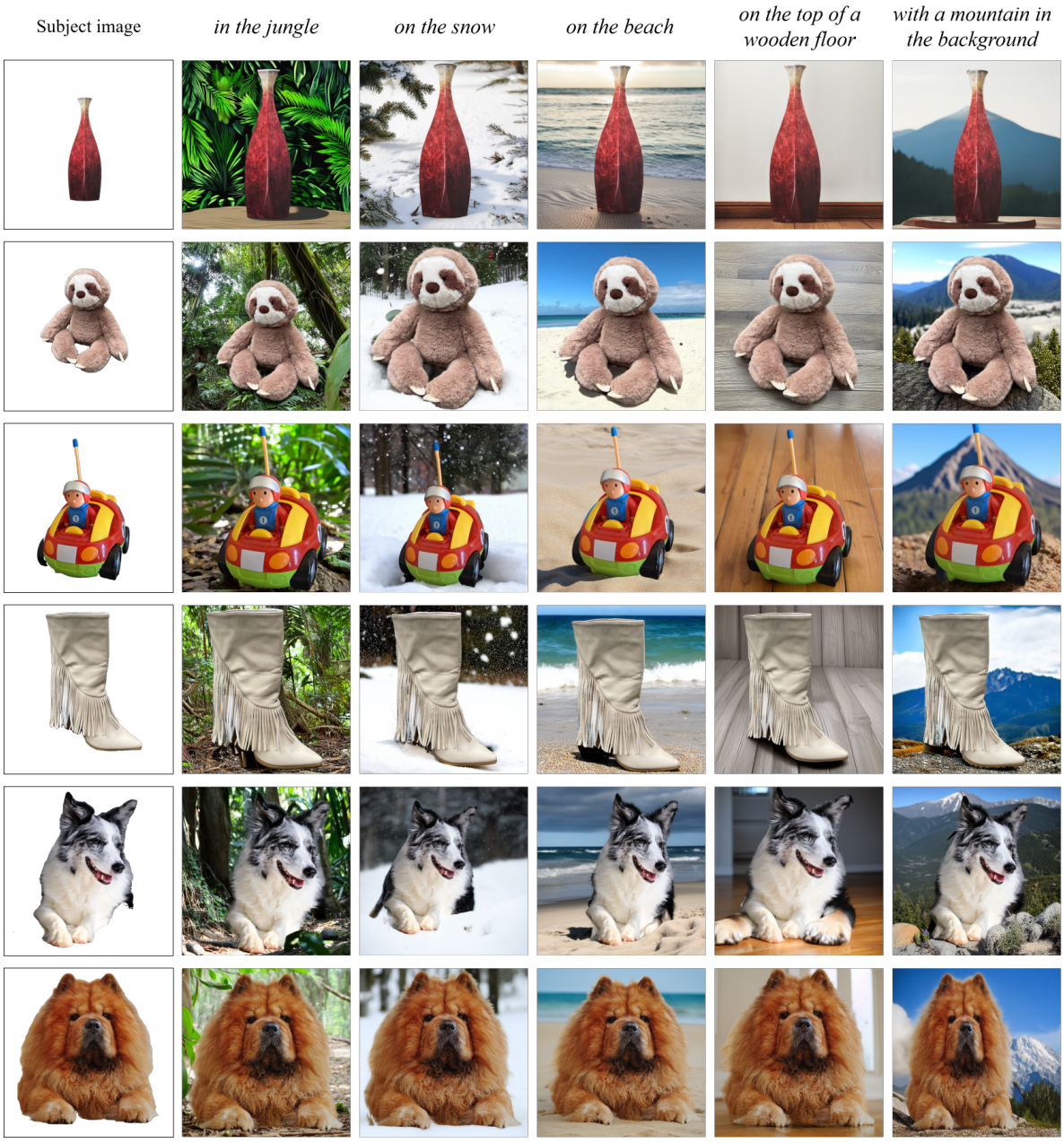


Figure 6. Visualization results of our UNIC-Adapter on DreamBench for subject-driven generation. The first column displays the subject images, while the other columns show the generated images based on different prompts.

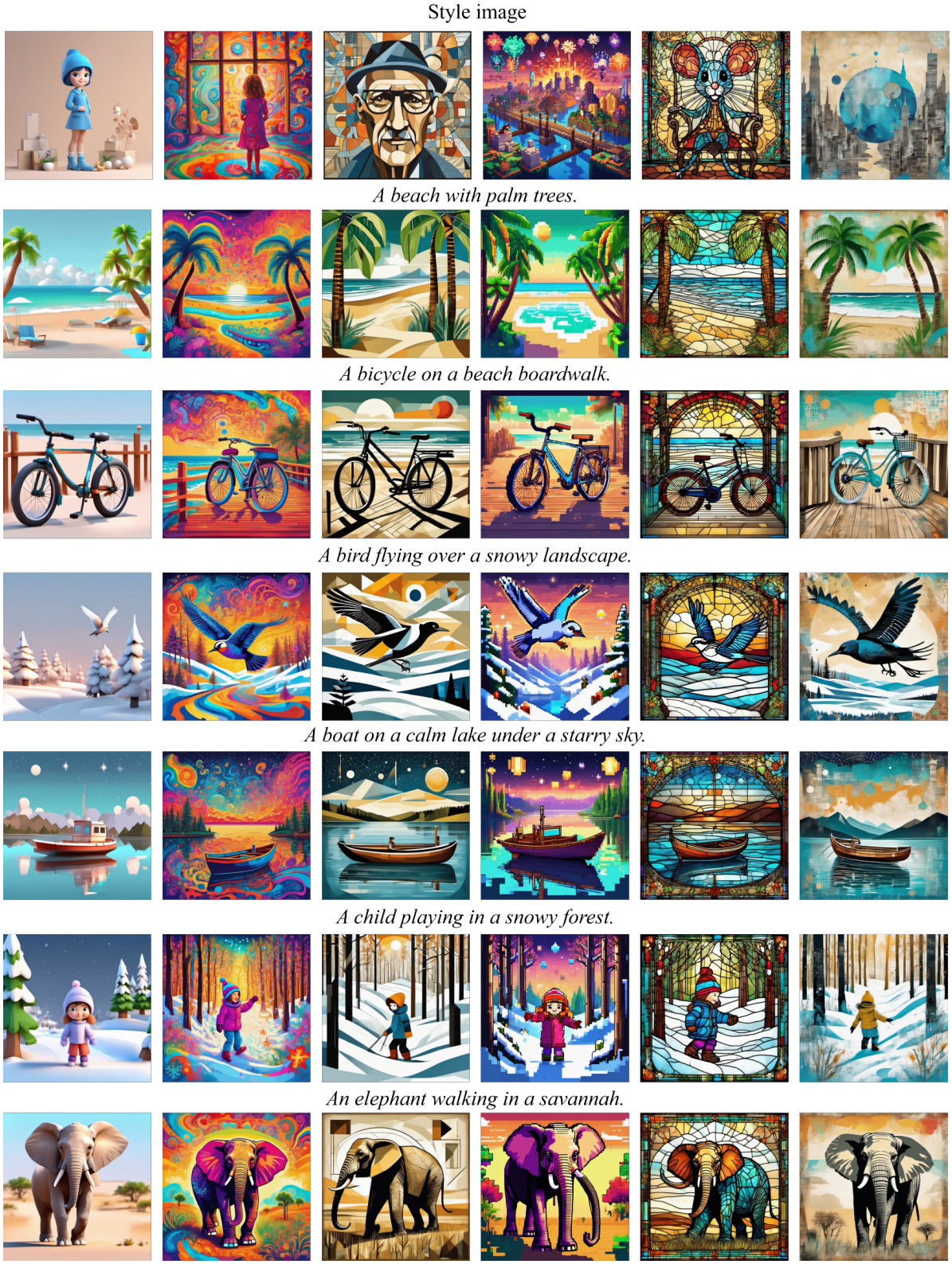


Figure 7. Visualization results of our UNIC-Adapter on style-image-based T2I generation. The first row shows the reference style images, and each subsequent row contains images generated from the same prompt, influenced by different style images.

References

[1] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 1

[2] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Int. Conf. Mach. and Learn.*, 2024. 1, 2

[3] Black Forest Labs. Flux: Inference repository. https://github.com/black-forest-labs/flux, 2024. Accessed: 2024-11-14. 2

[4] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-G: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*, 2023. 2

[5] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. UniControl: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023. 1

[6] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 3, 2024. 2

[7] Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. In *Int. Conf. Learn. Represent.*, 2025. 2

[8] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. 2