AnyControl: Create Your Artwork with Versatile Control on Text-to-Image Generation

Yanan Sun[1], Yanchen Liu[1,2], Yinhao Tang[1], Wenjie Pei[2], and Kai Chen[1⋆]

[1] Shanghai AI Laboratory
[2] Harbin Institute of Technology, Shenzhen
now.syn@gmail.com
wenjiecoder@outlook.com
{liuyanchen, tangyinhao, chenkai}@pjlab.org.cn

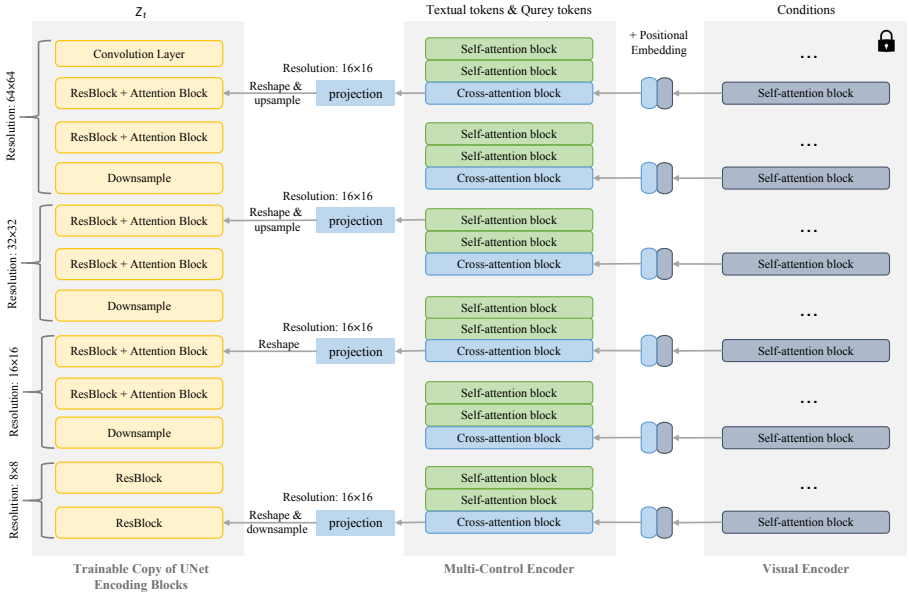

Fig. 1: Details of AnyControl and Multi-Control Encoder.

1 Implementation Details

Network. The detailed structure of our AnyControl is depicted in Figure 1. We base Stable Diffusion of version 1.5 to build our AnyControl. Similar to ControlNet [3], we make a trainable copy of the UNet encoding blocks for adapting to controlling information while freezing the pre-trained weights of Stable Diffusion model totally. In our Multi-Control Encoder, the number of query tokens is set to 256 enabling detailed controllable information extraction. The additional position embedding, with the same length as the query tokens, are shared by all input spatial conditions. We take the pre-trained weights of Q-Former [1] as the initialization for Multi-Control Encoder except for the query tokens and the additional position embedding, which are randomly initialized.

⋆ Corresponding author.

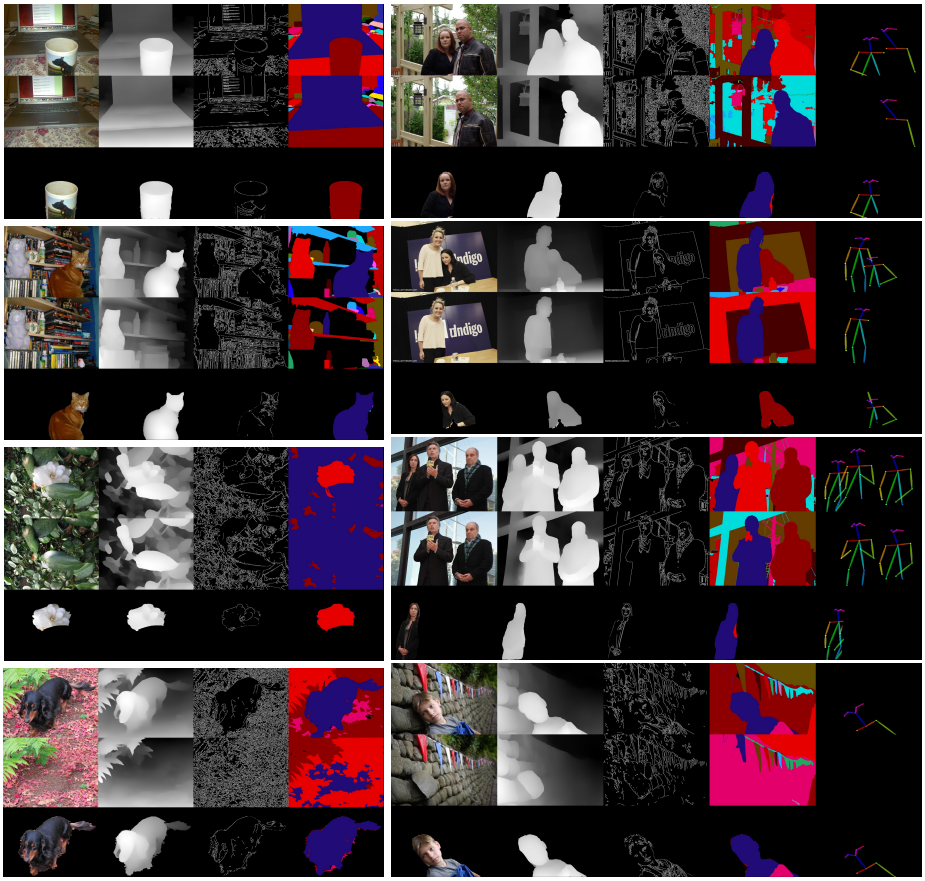

Fig. 2: More visualizations of unaligned data.

Hyper Parameters. We train AnyControl on 8 A100 GPU cards with a batch size of 8 on each GPU. We train the model for totally 90K iterations with a initial learning rate of 1e-5. During inference, we set the classifier-free guidance scale to 7.5. In all the experiments, we adopt DDIM [2] sampler with 50 timesteps for all the compared methods.

2 Unaligned Data

During producing the synthetic unaligned dataset, we utilize the groudtruth object masks with the area ratio in $[0.1, 0.4]$ to outline the foreground object, while oversmall or overlarge objects will lead to undesired recovered background image. PowerPaint [4] is a multi-task inpainting model supporting text-guided object inpainting, context-aware image inpainting as well as object removal. Here, we adopt the "object removal" mode for the unaligned data construction. More visualizations for synthetic unaligned data are in Figure 2.

3 AnyControl vs. Multi-ControlNet

As shown in Figure 3, multi-control methods with hand-crafted weights, *i.e.*, Multi-ControlNet [3], usually require a series of laborious weight adjustments according to the synthesized results while ours can automatically infer the combination weights and extracts unified multi-control embedding, thus producing harmonious results.

Fig. 3: Given prompt "cartoon style, a car parking in the canyon, a lion walking pass the car" and the edge conditions for a car as well as a lion, **left** shows hand-crafted weights adjustment for Multi-ControlNet [3]. X-axis and Y-axis represents the weight for lion and car conditions respectively. **Right** shows the results of AnyControl from three random seeds.

4 Multi-level Visual Tokens

Although the visual tokens from the last transformer block of the pre-trained visual encoder have already aggregated rich information, they are not sufficient to convey fine-grained controllable information. We conduct ablation experiments on the levels of used visual tokens from the visual encoder to the multi-control encoder. Table 1 demonstrate that integrating more visual tokens from middle layers increase FID and encounter performance saturation at 4-th level.

Table 1: Multi-level visual tokens. We gradually enable the visual tokens from the deepest level to the shallowest level.

Levels	1	2	3	4	5	6
FID ↓	45.64	43.73	43.69	43.67	43.74	44.28
CLIP ↑	26.35	26.40	26.39	26.39	26.38	26.40

5 More Qualitative Results

More qualitative results on multi-control synthesis are shown in Figure 4. Results of single-control synthesis including depth map, edge map, segmentation map and human pose are shown in Figure 5 to Figure 8 respectively.

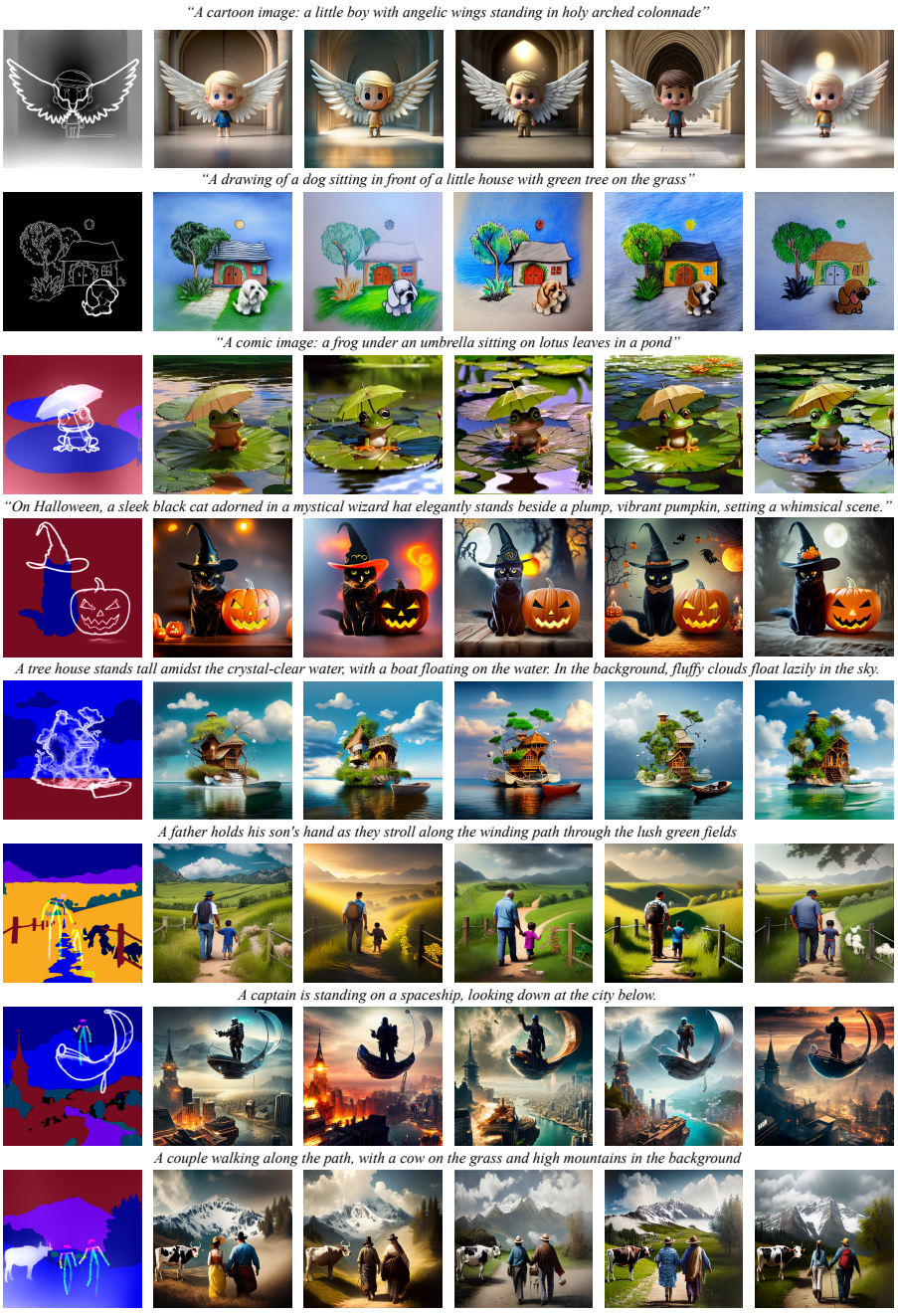

Fig. 4: More visual results from AnyControl on multi-control image synthesis.

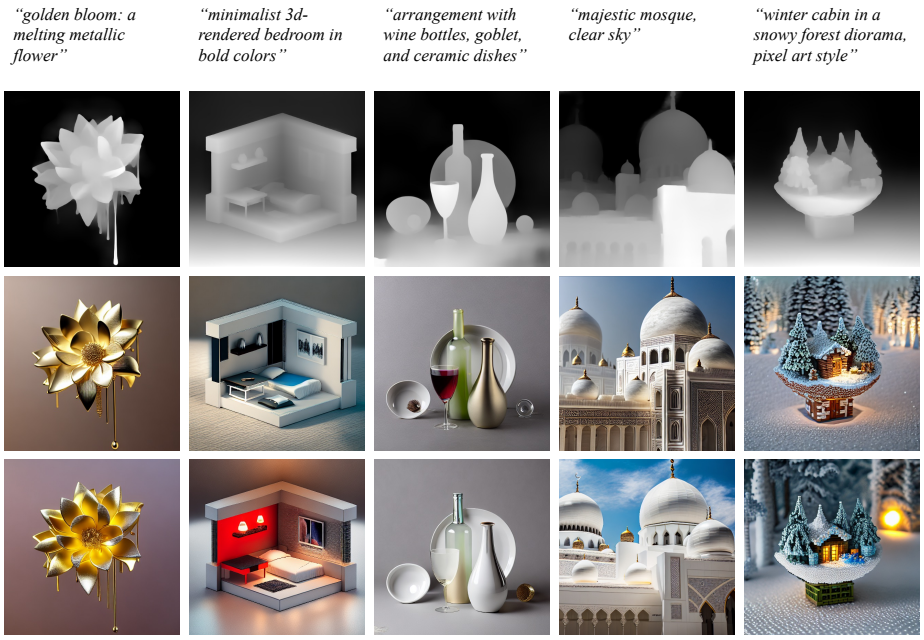

Fig. 5: Visual results on depth controlled image synthesis.

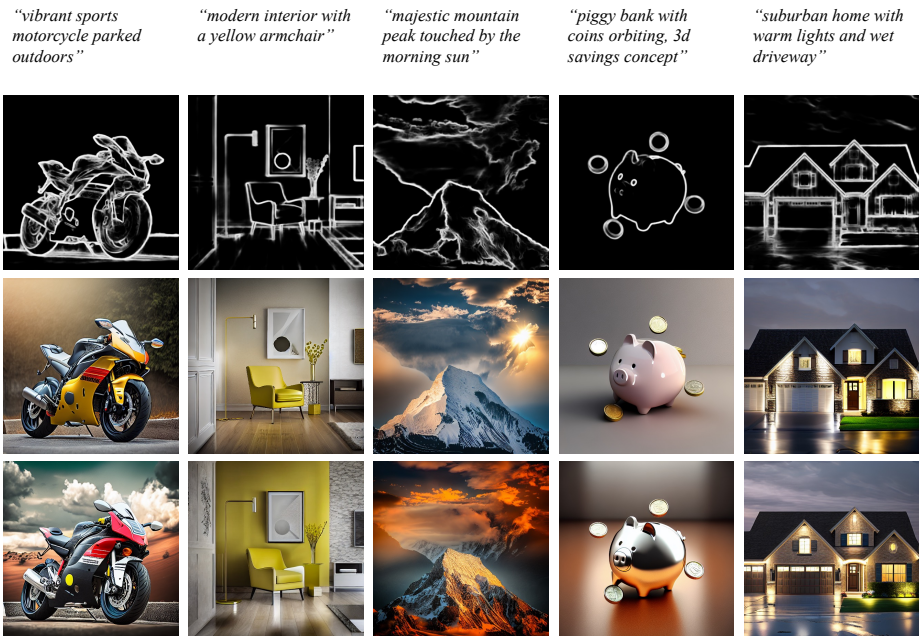

Fig. 6: Visual results on edge controlled image synthesis.

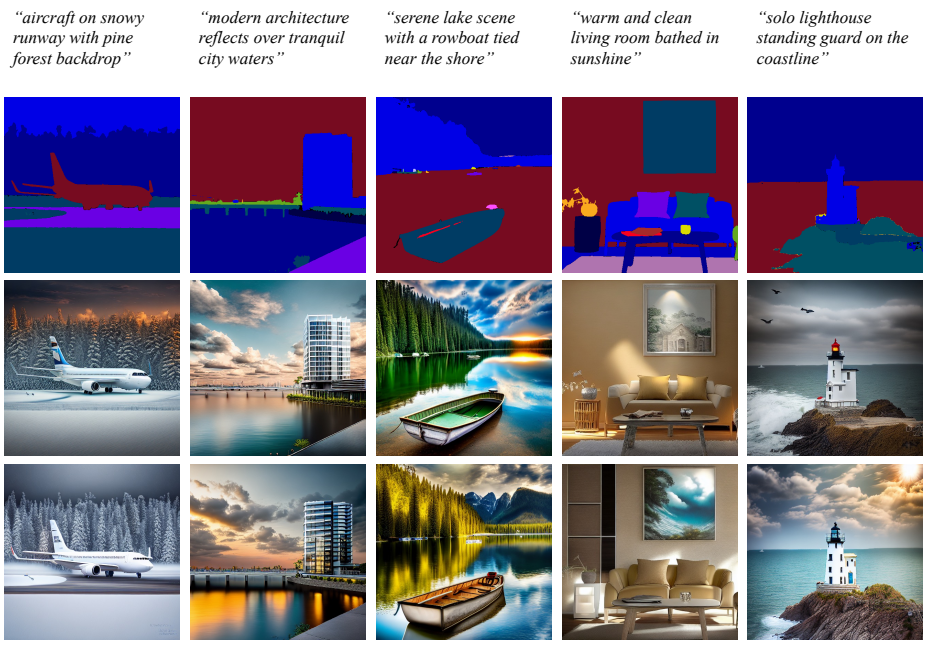

Fig. 7: Visual results on segmentation controlled image synthesis.

Fig. 8: Visual results on human pose controlled image synthesis.

References

1. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023) 1
2. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 2
3. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023) 1, 3
4. Zhuang, J., Zeng, Y., Liu, W., Yuan, C., Chen, K.: A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. arXiv preprint arXiv:2312.03594 (2023) 2