

Supplementary Materials



Figure 10: Exemplary training image pairs from the source datasets.

Number of loops	CLIP-T	CLIP-I	DINO
0	0.321	0.714	0.560
1	0.317	0.732	0.627
5	0.316	0.777	0.694
10	0.308	0.791	0.720
Auto-stop (avg. 6.67)	0.316	0.782	0.718

Table 6: Evaluating subject-driven image generation task on DreamBench dataset with different iterative loops. The performance of the auto-stop mechanism reaches the best balance between subject identity and text following, resulting in 6.67 average steps for each evaluation pair on the DreamBench dataset.

A DATASET CONSTRUCTION

The training dataset is constructed using two widely recognized datasets: COCO2014(Lin et al. (2014)) and YoutubeVIS(Yang et al. (2021)), examples are illustrated in Fig. 10. **COCO2014 Dataset.** We crop 1 to 4 objects from a given target image to serve as the subject images. Each cropped object, along with the corresponding full image, forms a subject-target training pair. This pairing ensures that the model learns the association between individual subjects and the broader scene in which they are located.

YoutubeVIS Dataset. The YoutubeVIS dataset contains videos with annotated instances of objects over time. To create training pairs, we extract images of the same subject from different frames of a video, following the methodology proposed in Chen et al. (2024). This process captures the variations in appearance, pose, and position of the same subject across different frames, providing valuable temporal data that helps the model learn consistent subject identification even in dynamic scenes.

B AUTOSTOP MACHANISM

We use the DINOv2 image encoder as the criteria calculator. For each newly generated image and the image from the previous loop, we calculate patch-wise similarity using the DINOv2 encoder. If the similarity exceeds a predefined threshold, the iterative process stops, indicating sufficient subject feature transfer. To handle cases where the layout image differs significantly from the subject image, leading to weak injection, we enforce a minimum of 3 loops. For difficult cases where feature map similarity remains low, we cap the maximum loop count at 10. We evaluate performance on the DreamBench dataset with varying loop numbers in Tab. 6.

C MORE VISUALIZATION RESULTS

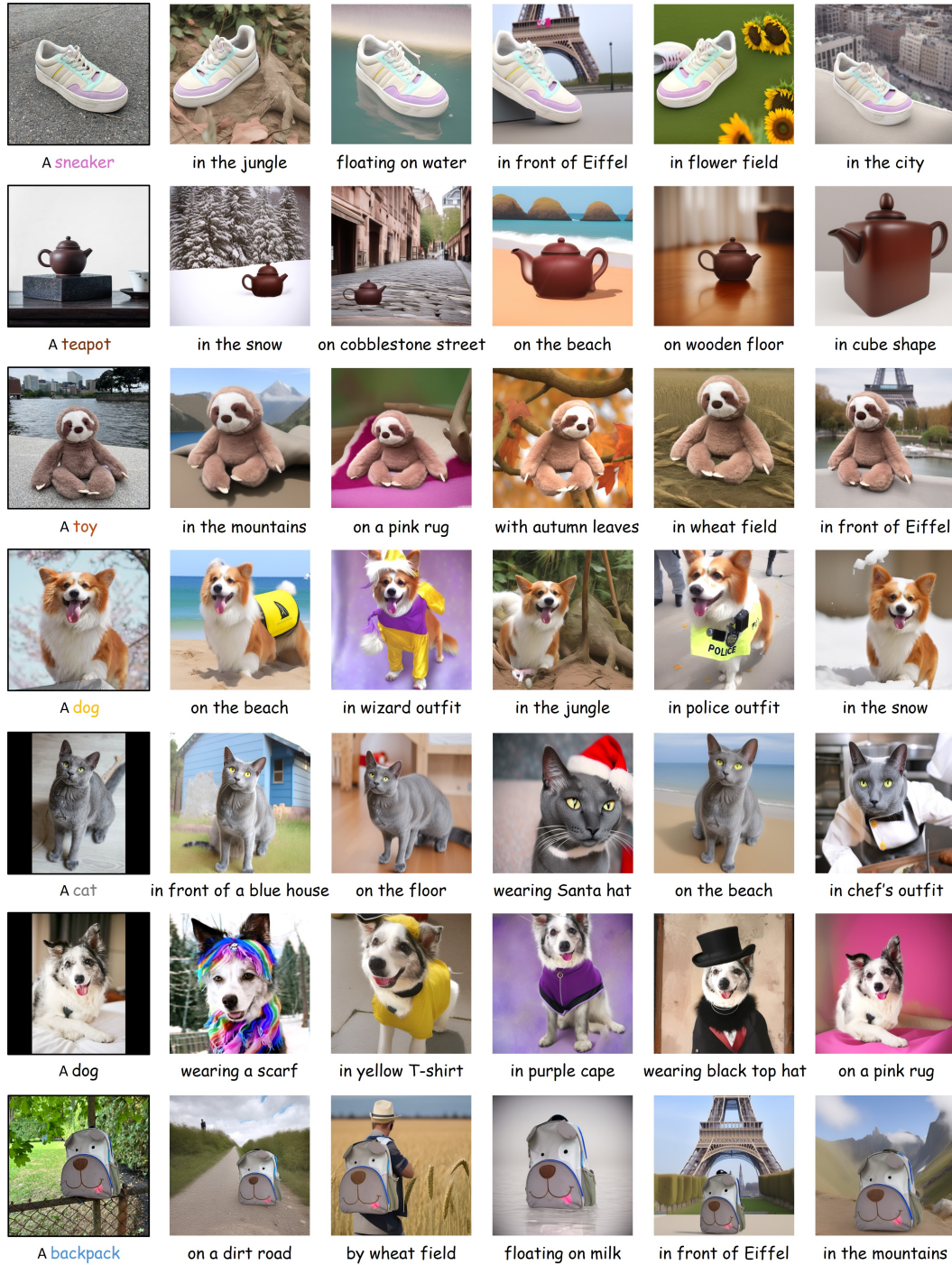


Figure 11: More visualization results for subject-driven image generation.

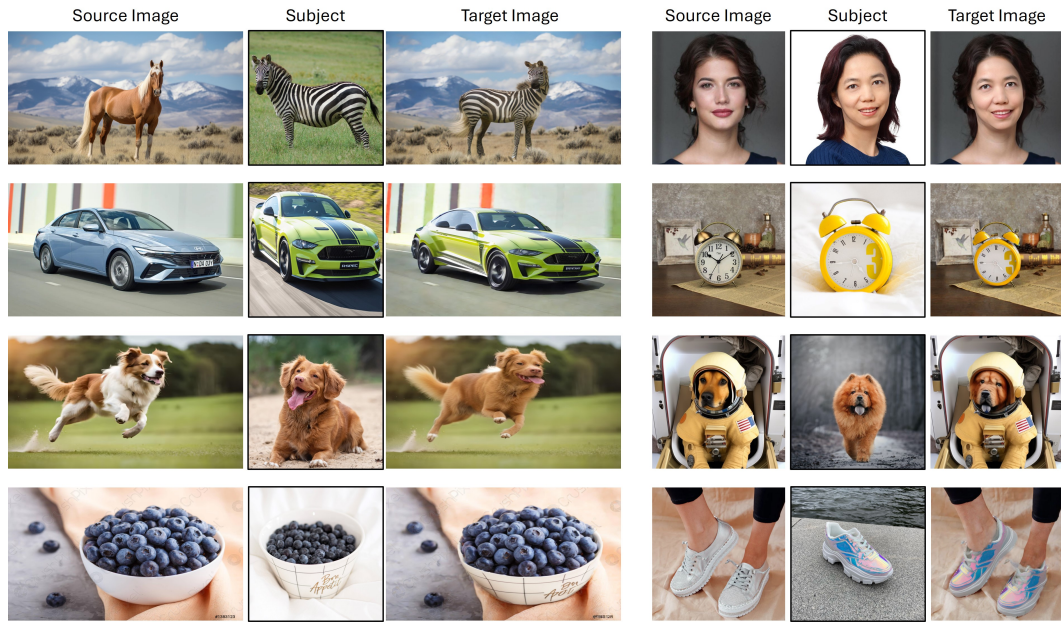


Figure 12: More visualization results for subject-driven image editing.



Figure 13: More visualization results for human content generation.

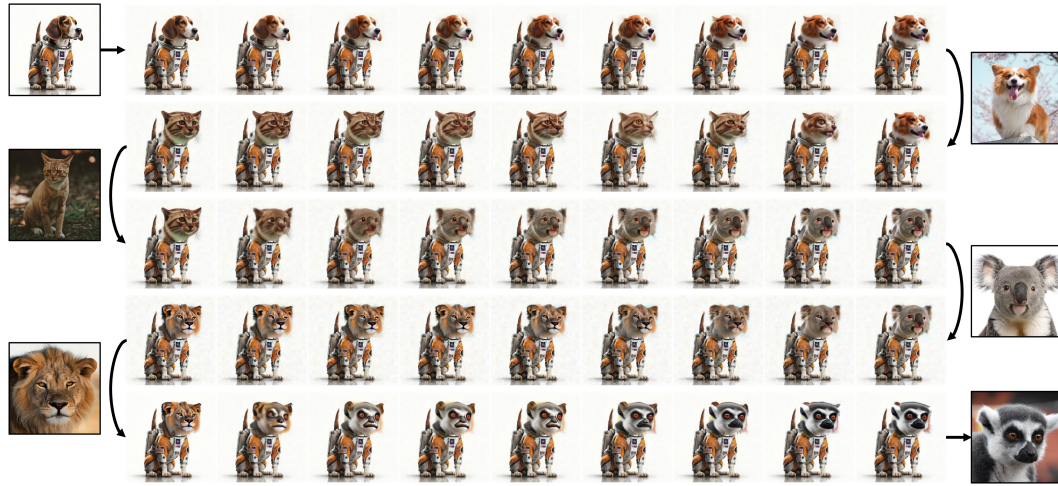


Figure 14: Interpolation between subjects.