

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.

756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809

C MORE VISUALIZATION RESULTS

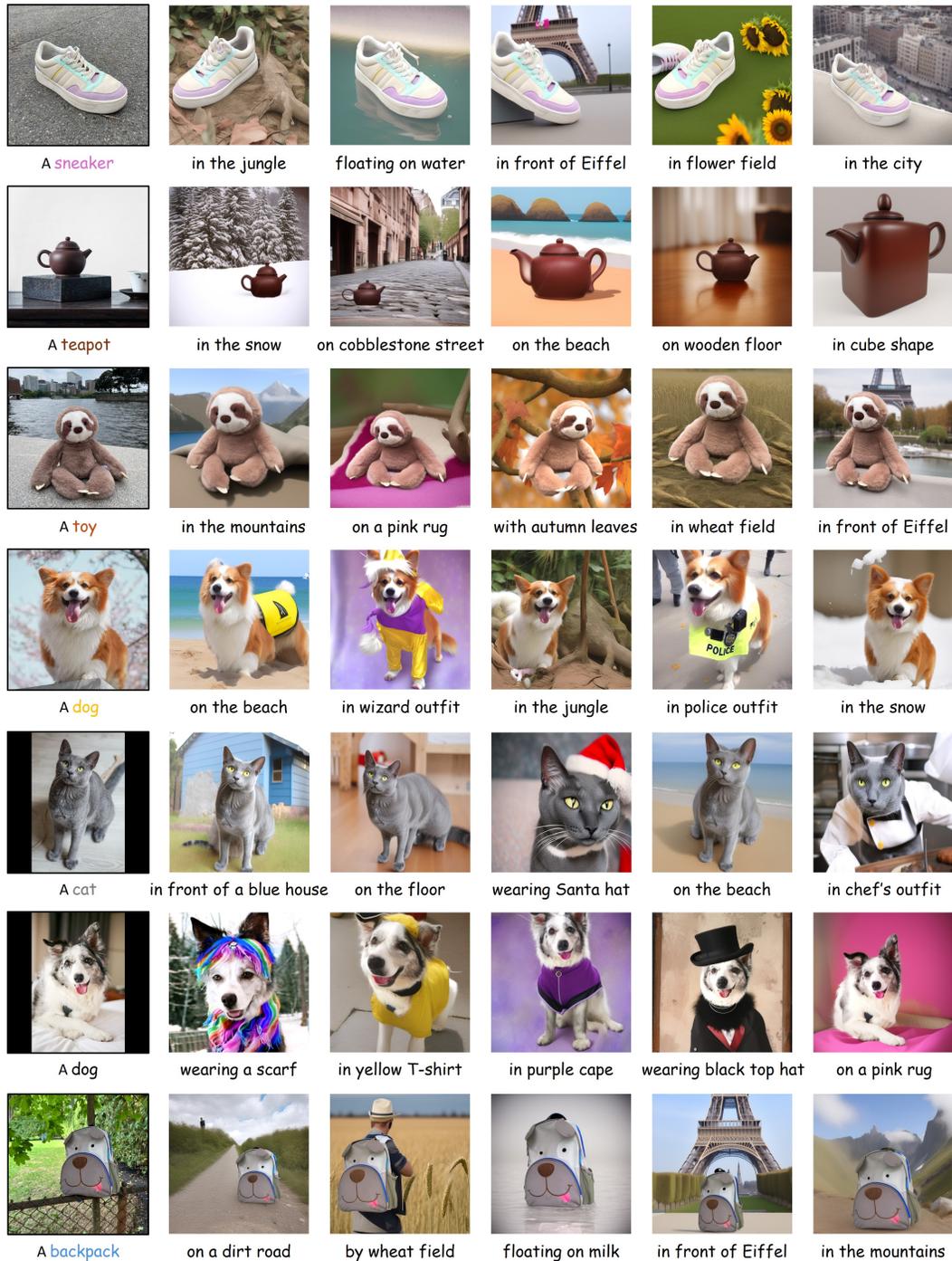


Figure 11: More visualization results for subject-driven image generation.

810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863

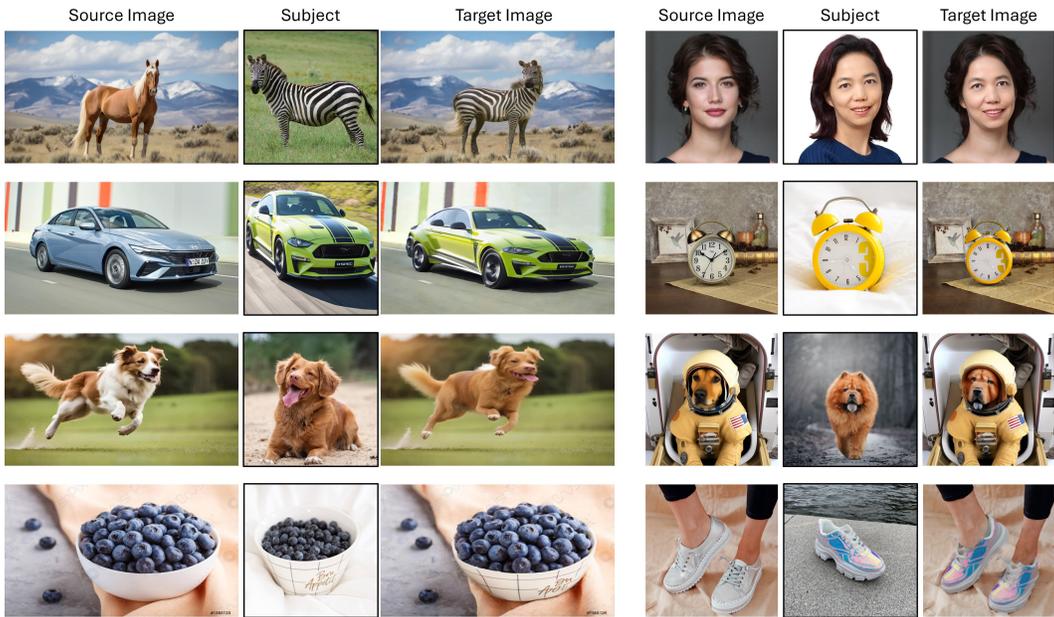


Figure 12: More visualization results for subject-driven image editing.



Figure 13: More visualization results for human content generation.

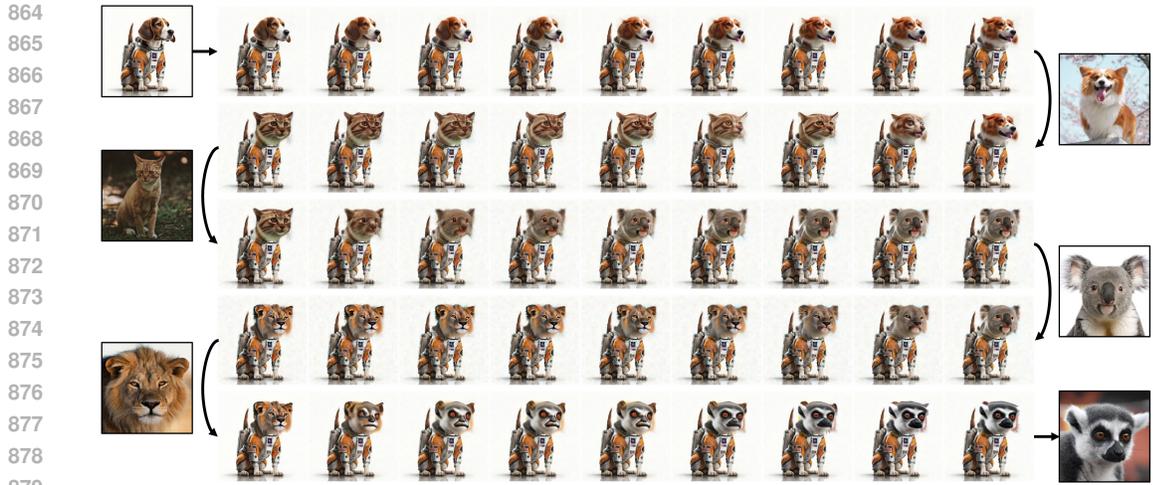


Figure 14: Interpolation between subjects.