

Supplementary Materials: DisenStudio

Anonymous Authors

In the appendix, we provide the qualitative results of the SDCA ablation study and more visual comparisons between DisenStudio and baselines. Additionally, we also provide the detailed diagram of the spatial-disentangled cross-attention of each figure used in our main manuscript. Finally, we will discuss the limitations of our work and potential future directions.

R1 QUALITATIVE RESULTS ABOUT SDCA EFFECTIVENESS

In our main manuscript, we provide the quantitative results of the SDCA, and here we provide qualitative analysis. In Figure R1, we compare the results of DisenStudio, DisenStudio w/o SDCA, our competitive baseline VideoDreamer, and VideoDreamer+SDCA. From the results, we can see that without SDCA, the model fails to assign the red hat and the yellow scarf to the right subjects. Additionally, the attribute of the cat w/o SDCA is changed. The dog of VideoDreamer seems to be mixed with the feature of the cat, and it also fails to assign the hat and scarf to the respective subjects. When we apply SDCA to VideoDreamer, the hat and the scarf can be appropriately assigned, but the visual attributes of the cat and the dog are not well preserved across frames, indicating the necessity of our fine-tuning strategy. Additionally, VideoDreamer and VideoDreamer+SDCA fail to generate the swimming pool background, indicating its overfitting to the finetuning images.

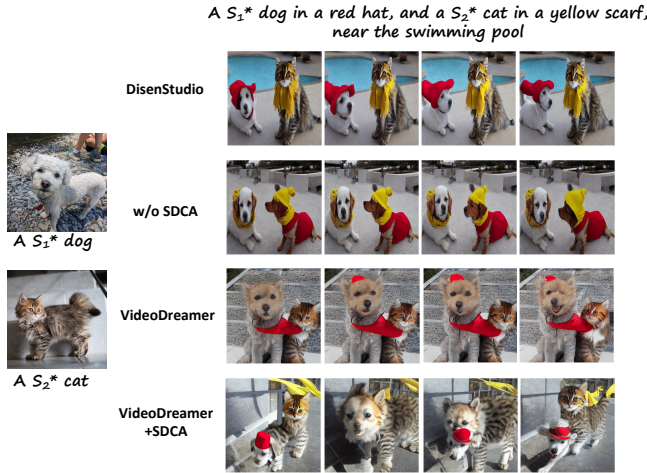


Figure R1: Comparison among Disenstudio, w/o SDCA, VideoDreamer and VideoDreamer+SDCA.

R2 MORE QUALITATIVE COMPARISON

We provide more qualitative comparisons with baselines in Figure R2. The observations are consistent with the results in our main manuscript, where the baselines suffer from subject-missing, attribute-binding, and action-binding problems. Our proposed DisenStudio outperforms them clearly.

R3 DETAILS OF SDCA

We provide the detailed spatial-disentangled cross-attention of each figure used in our main manuscript. The detailed diagram is shown in Figure R3, where for each presented figure (Figure 6 to Figure 10) in the main manuscript, we show its corresponding disentangled-spatial cross-attention, where we present the prompt used for cross-attention and the regions it attends to. The prompt and region are matched in color, e.g., red-color prompts will work on red regions and blue-color prompts will work on blue regions. Particularly, white regions are matched with black-color prompts, which indicate the background of the videos, e.g., “on the beach, on the grass”.

R4 LIMITATIONS AND FUTURE WORKS

In this paper, we propose a DisenStudio framework for customized multi-subject text-to-video generation. Despite its significant superiority over existing methods, it still belongs to the prior works in this field and has several limitations. First, we adopt AnimateDiff [1] as our base model, we inherit its limitations, where AnimateDiff can only generate 16-frame videos and fail to generate longer videos. Due to its limited video length, it fails to generate videos where the scenario and the motions of the subject have large changes, such as “two girls first dance in the gym and then walk to the swimming pool, and finally swim in the pool”. Therefore, one future direction is how to adapt our work to a more advanced base model that can generate longer videos. Additionally, the motions of multiple subjects are from the base model, future works can consider how to customize a particular motion for each subject. Finally, due to the resolution of the AnimateDiff being fixed to 512, when customizing more subjects, each subject can only cover fewer pixels, making each subject lose some visual details. Future works can also focus on solving the resolution problem to support more subject customization.

REFERENCES

- [1] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. 2023. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725* (2023).

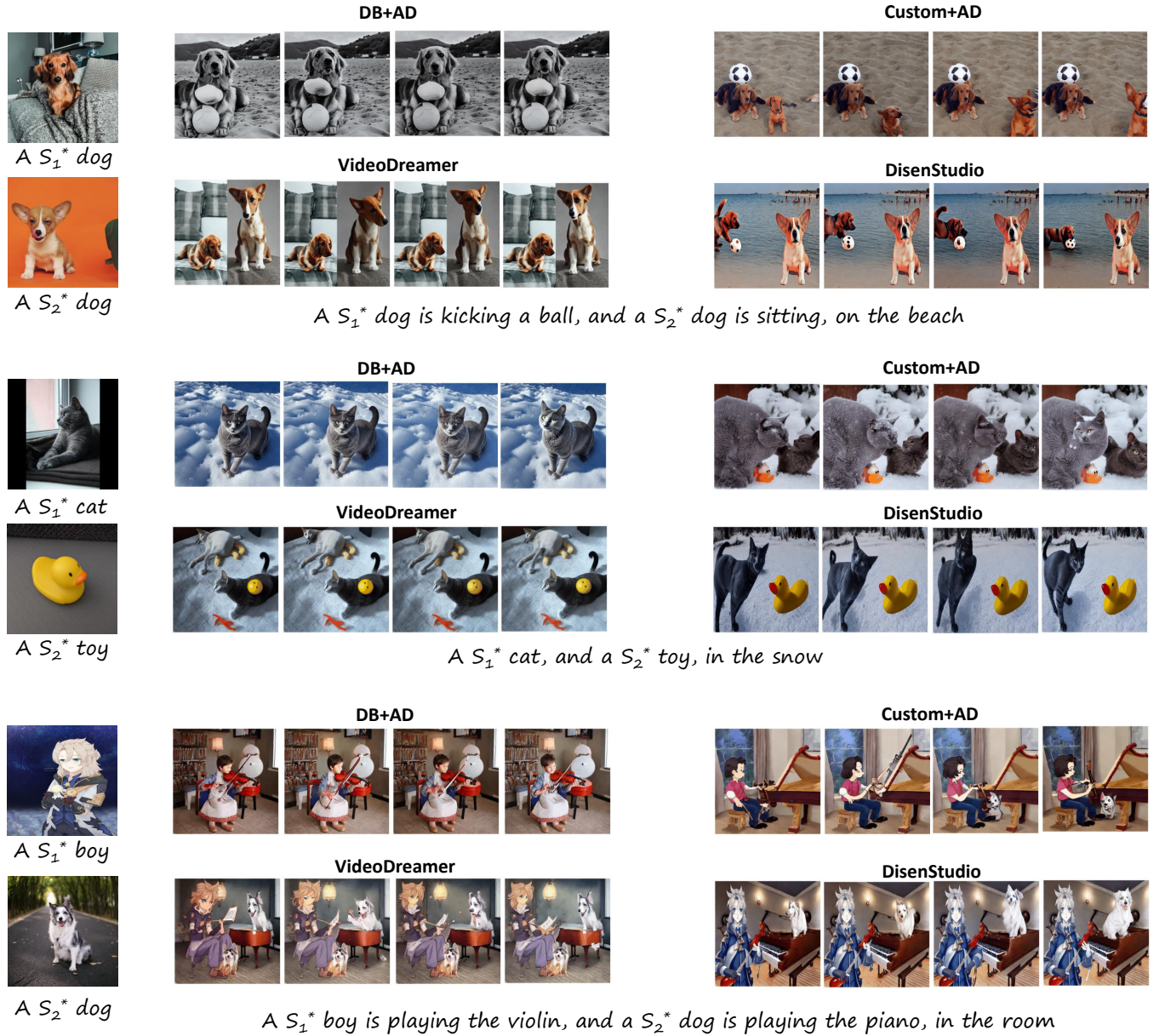


Figure R2: Qualitative comparison between DisenStudio and baselines. DreamBooth+AnimateDiff(DB+AD) often suffers from missing one subject. CustomDiffusion+AnimateDiff(Custom+AD) suffers from attribute-binding problem, where it often generates two similar subjects instead of the two given subjects. VideoDreamer also suffers from attribute-binding and action-binding problems. In contrast, our proposed method, DisenStudio, best preserves the visual details of each subject, and also assigns the right action to each subject.

Figure 6



Figure 7

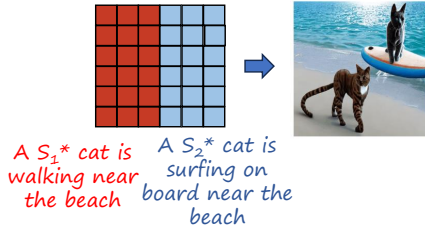


Figure 8

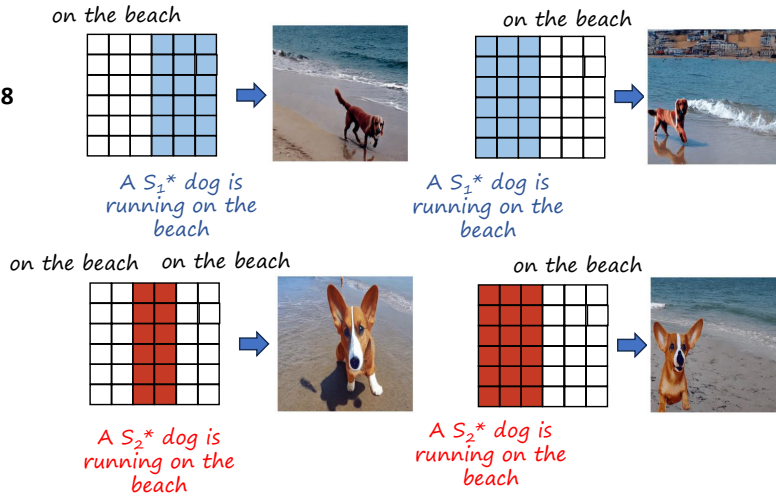


Figure 9

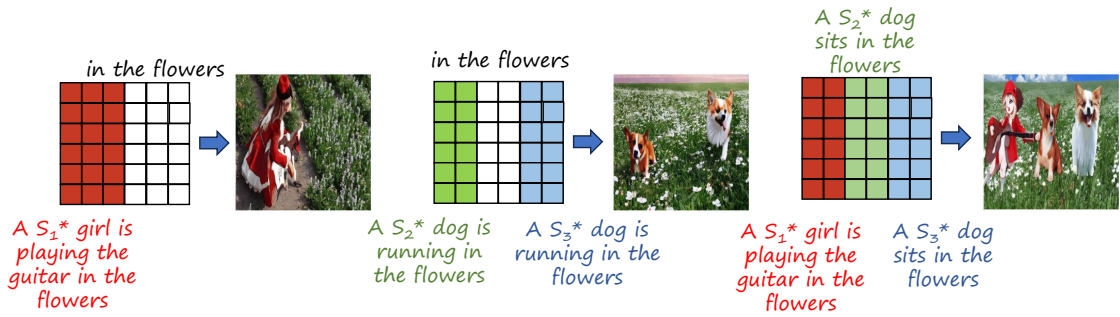


Figure 10



Figure R3: The diagram of the spatial-disentangled cross-attention used for the figures in the main manuscript. The prompts and the regions are matched in color, and particularly, white regions are matched with black-color prompts which indicate the background of the video.