Supplementary Materials: CustomCrafter: Customized Video Generation with Preserving Motion and Concept Composition Abilities

Anonymous submission

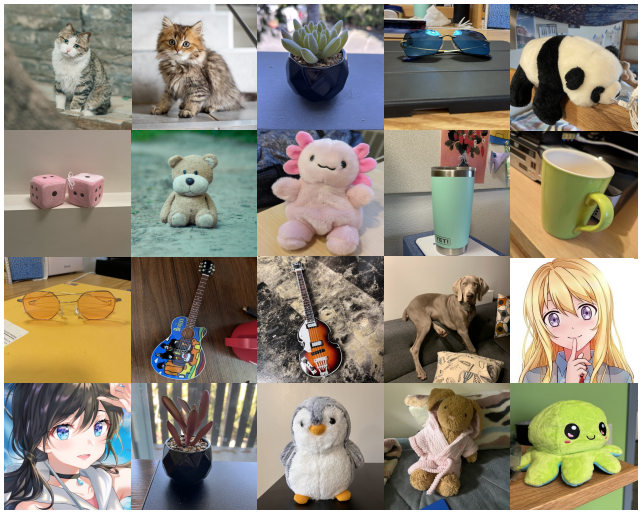

Figure 1: The overview of our experimental dataset. Our experimental dataset cover a wide range of object categories.

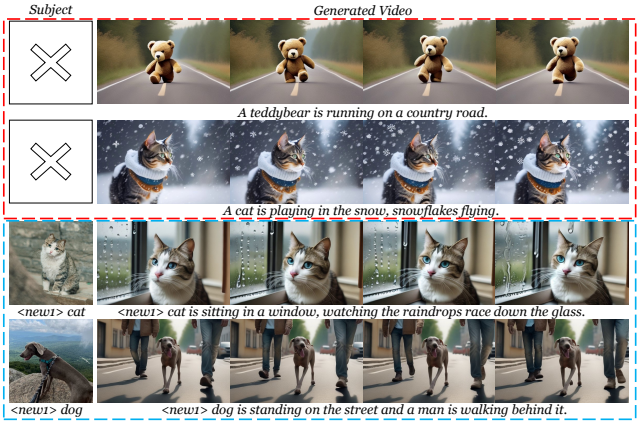

Figure 2: Limitations. The first two rows are the videos generated by VideoCrafter2 without using our method for customized generation training. The last two rows are the videos generated by our method after optimizing the prompt.

Experimental Details

In this section, we have provided additional experimental details. We have detailed the data used in our experiments, as well as the hyperparameter settings and replication methods for the two baseline methods we reproduced. All experiments are conducted using VideoCrafter2 as the base model and using 4 NVIDIA A100 cards. During training, we randomly resize the target images from 0.4 - 1.4× and append the prompt "very small", "far away" or "zoomed in", "close up" accordingly to the prompt based on the resize ratio.

Datasets

We select subjects from DreamBooth (Ruiz et al. 2023), Custom Diffusion (Kumari et al. 2023) and Mix-of-Show (Gu et al. 2024) for a total of 20 customized subjects. This dataset includes various types of subjects, such as dolls, sunglasses, guitars, water cups, and anime characters, to better evaluate the performance of custom video generation methods on different types of items, and to verify the universality of the method. The overall content of the dataset is shown in Figure 1.

Custom Diffusion

We have reproduced Custom Diffusion as a baseline on the base model of VideoCrafter2 (Chen et al. 2024). We follow the method mentioned in the paper (Kumari et al. 2023), where we only update the key and value of spatial-cross-attention during training. The hyperparameters used are: training for 2000 steps, setting the batch size to 1, and the learning rate to 1×10^{-5}. In addition, for the CustomDiffusion* mentioned in the Method section of the main paper, we further extended the training steps to 10000 to ensure the training time is the same as our method.

DreamVideo

For DreamVideo (Wei et al. 2023), it introduces additional video data during the motion learning process and fine-tunes it to recover the model's generation ability. However, our method does not introduce additional data for fine-tuning. Therefore, for a fair comparison, we only reproduced its subject learning part for comparison. Following (Wei et al. 2023), we reproduced the Subject Learning part of DreamVideo on the base model of VideoCrafter2 (Chen et al. 2024) by adding an identity adapter to learn the subject's appearance features. We take 3000 iterations to optimize the textual identity with a learning rate of 1×10^{-4} and 800 iterations to learn the identity adapter with a learning

rate of 1×10^{-5}. In addition, for DreamVideo* mentioned in the Method section of the main text, we further extended the training steps of the identity adapter to 3000 to allow the identity adapter to better capture the appearance features of the subject.

Technical Appendix Video

To better and more intuitively demonstrate the effectiveness of our method, we have created a demonstration video. We strongly recommend that the reader experience the effect of our method more intuitively through the video. In the video, it is clear that our method has better subject appearance consistency, concept combination ability, and motion generation ability compared to other methods without using additional video data for training to recover motion generation ability. After only subject learning training, our method can retain the model's original concept combination and motion generation abilities to generate videos of the specified subject.

More Visualizations for Performance Comparison

In this chapter, we present more results of our method and comparison methods to demonstrate the improvement of our method in motion generation, concept combination, and subject appearance consistency.

As shown in Figure 3, our method is stronger than existing methods in terms of conceptual combination ability and motion generation ability. For instance, in Figure 3(a), the task is to create a video of a teddy bear running on a country road. While other methods overfit the "frozen" motion in the training data, thereby failing to produce a running motion, our method adeptly generated a video that aligns with the prompt. In Figure 3(b), when we want to generate a video of a panda doll sitting on a windowsill, other methods cannot generate this conceptual combination, but our method can generate a video that matches the prompt description. Moving to Figure 3(c) and (d), while our method achieves the best effect in ensuring the consistency of the cat's appearance, the realism of the interaction between the cat and the environment in the generated video and the degree of fit with the prompt description are significantly better than the other two methods. In Figure 3(e), when we want to generate a scene of a plush dice toy rotating and falling in the sky, our method also generates more obvious motions than other methods. Other methods overfit the training data, do not recombine objects with other concepts, and have a smaller range of motions. Figure 3(f) illustrates another instance where our method shined, this time by generating a video of an person playing our designated guitar. While other methods cannot normally generate such a conceptual combination, our approach not only generates the conceptual combination of people and guitars normally, but also has a significant improvement in motion generation ability. Moreover, as shown in Figure 3(h), when we want to generate a video of a dog walking on the street, our method also generates motions closer to reality. Besides, our method can generate multiple learned subject objects, as shown in Figure 3(g). Only our method correctly generates two water cups and the "close-up shot" required by the prompt, which shows that our method can better utilize the capabilities of the video generation model to complete customized video generation.

In addition, we used static prompts to generate videos, focusing on comparing the consistency of the subject's appearance and the concept combination ability of our method. As shown in Figure 4, our method has made significant progress in capturing the details of the subject's appearance compared to existing methods. This further demonstrates the superior performance of our approach in maintaining the fidelity of the subject's appearance and combining different concepts effectively. As shown in Figure 4, when we want to generate a video of a child sleeping with a specified dice toy, other methods cannot correctly generate the concept combination of the child and the dice toy, resulting in a poor overall generation effect. At the same time, when we want to generate a scene where the specified potted plant is surrounded by other potted plants, we find that other methods cannot integrate the given potted plant well into the scene. However, our method can normally generate such static videos, indicating that our method has a good ability to combine concepts. The mastery of the subject's appearance details can also be seen from static videos. For example, as shown in the bottom right corner of Figure 4, our method captures the characteristic of the glasses with more accuracy compared to other methods. At the same time, our method successfully connects the three concepts of seashore, sandy beach and specific sunglasses perfectly, generating a video that matches the prompt perfectly. Other methods have not achieved this. In addition, as shown in the left right corner of Figure 4, our method can also accurately capture the given cat's appearance, generating higher quality videos. This further demonstrates the superior performance of our method in maintaining the fidelity of the subject's appearance and preserving the concept composition abilities.

Limitations

We found that both our method and other methods had the problem of unclear generated background, so we conducted research on this issue. The blurred background is caused by VideoCrafter2 (Chen et al. 2024), which our method is based on. As shown in the red box in Figure 2, we show the results of the original VideoCrafter2 using the same prompt as in Figure 1 in the paper. The videos generated by VideoCrafter2 also have the same phenomenon that the subject is clear and the rest of the background is blurred. Therefore, we believe that this is a characteristic of VideoCrafter2. In addition, as shown in the blue box in Figure 2, we find that adding specific background descriptions to the prompt will help generate background details. For example, we can see the background details of "raindrops race down the glass" and "a man is walking behind it". Therefore, the quality of the videos generated by our method is constrained by the performance of the base model. So, we believe that with the improvement of the underlying model's performance, our method will be capable of producing better and higher-quality videos.

Figure 3: Qualitative comparison of customized video generation with both subjects and motions. Without guidance from additional videos, our method significantly outperforms in terms of concept combination and motions.In addition, the subject appearance consistency of our method is significantly better than that of existing methods on multiple samples.

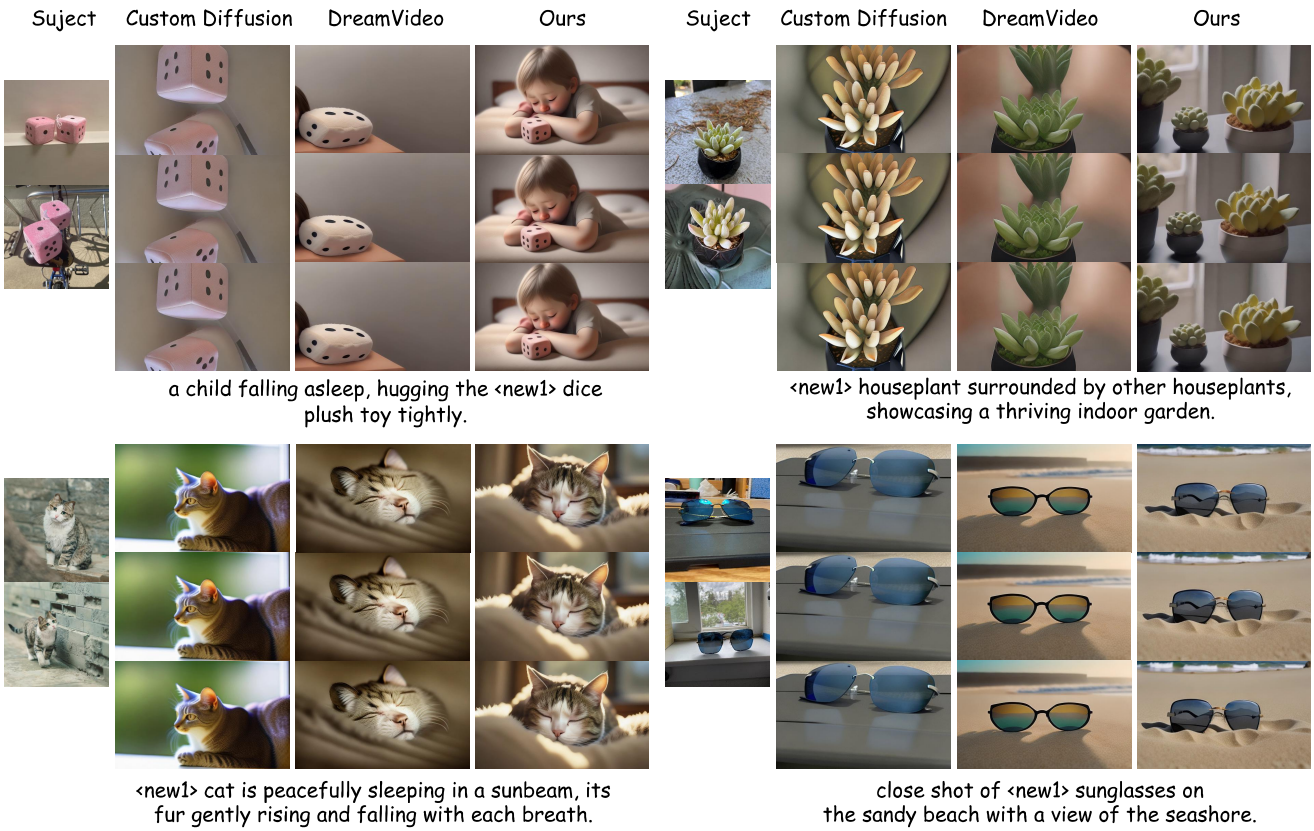

Figure 4: Qualitative comparison of customized video generation use the static prompt generated video for comparison. It can be seen that our method can better learn the details of sunjects during the appearance learning process, and has better concept combination capabilities.

References

Chen, H.; Zhang, Y.; Cun, X.; Xia, M.; Wang, X.; Weng, C.; and Shan, Y. 2024. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*.

Gu, Y.; Wang, X.; Wu, J. Z.; Shi, Y.; Chen, Y.; Fan, Z.; Xiao, W.; Zhao, R.; Chang, S.; Wu, W.; et al. 2024. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *Proc. NeurIPS*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. CVPR*, 22500–22510.

Wei, Y.; Zhang, S.; Qing, Z.; Yuan, H.; Liu, Z.; Liu, Y.; Zhang, Y.; Zhou, J.; and Shan, H. 2023. Dreamvideo: Composing your dream videos with customized subject and motion. *arXiv*.