Frame by Familiar Frame: Understanding Replication in Video Diffusion Models

Aimon Rahman *, Malsha V. Perera *, and Vishal M. Patel
Johns Hopkins University
* denotes equal contribution
{arahma30,jperera4,vpatel36}@jhu.edu

1. Video Replication in SOTA video generation model

Video sample replication is a significant challenge in state-of-the-art models, especially when models and their training datasets are not publicly accessible. In our research, we typically analyze generated videos from project websites and compare them to the closest matches in their training data. This becomes more complicated when the training datasets themselves are not available. In this analysis, we focus on the VideoFusion [2] model, a recent state-of-the-art example where the generated videos are not accessible. To address this, we use screenshots from the model's research paper, representing the generated videos, and match them with the training dataset. Our findings, as illustrated in Figure 1, reveal that even the latest models are susceptible to replicating videos from their training data.

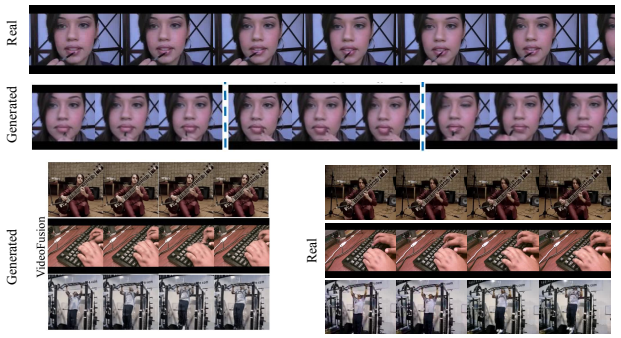

Figure 1. The highest similarities identified within the UCF-101 dataset compared against outputs generated by an unconditional video generation approach, utilizing VideoFusion [2]. The generated videos are sourced from the research paper.

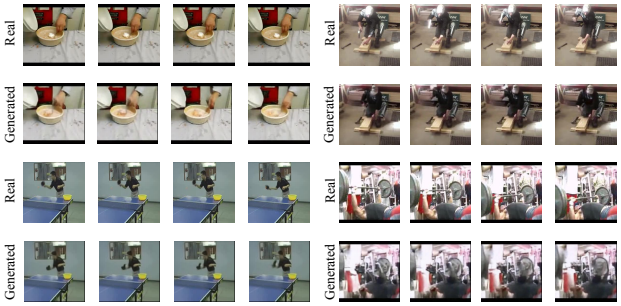

Figure 2. The highest similarities identified within the UCF-101 dataset compared against outputs generated by an unconditional video generation approach, utilizing LVDM [1]. The generated videos are sourced from the project website.

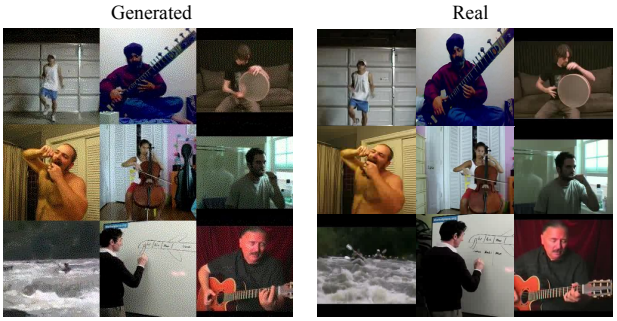

Figure 3. The highest similarities identified within the UCF-101 dataset compared against outputs generated by an unconditional video generation approach, utilizing VIDM [3]. The generated videos are sourced from the project website.

2. More Qualitative Examples of Video Replication

In this section, we present additional qualitative examples of video replication from various video diffusion models. We examine two distinct types of models: a general video diffusion model operating in the pixel domain [3] and a latent diffusion model [1]. Figures 3 and 2 provide examples of replication observed in both cases, showcasing the phenomenon across different model architectures.

2.1. Discussion on Replication in Video Diffusion Models

Video diffusion models demonstrate a higher susceptibility to replication compared to image diffusion models, making the originality of generated videos a relatively unexplored area. This raises important questions about the extent to which these models can produce original content. Replication tendencies are evident in both short and long-form videos generated by these models. We propose that if the generated videos lack realism (as seen in models like MakeAVideo [4], LVDM T2V [1], etc.), they are less likely to be replicas. This observation suggests a shift in the focus of current research in this field.

References

[1] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 1, 2

[2] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023. 1

[3] Kangfu Mei and Vishal Patel. Vidm: Video implicit diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9117–9125, 2023. 1

[4] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2