VideoAlign: A Comprehensive Model for Evaluating Alignment Between Text and Generated Videos

Yuanming Yang[1] **Xiaoqian Liu**[2] **Jian Chen**[2]
[1]Institute for Interdisciplinary Information Sciences, Tsinghua University
[2]Department of Automation, Tsinghua University
{yym22,lxq21,chenjian20}@mails.tsinghua.edu.cn

Abstract

Text-to-video generation models have made significant progress recently, but challenges remain in achieving alignment with human preferences. The generated videos frequently lack reliable consistency with their corresponding textual descriptions, and manual evaluation is both labor-intensive and expensive. This study proposes a comprehensive solution to address these alignment issues. We will introduce **VideoAlign**, an end-to-end reward model designed to automatically evaluate the instruction-following capabilities of video generation models.

1 Introduction

Text-to-video (T2V) technology, driven by generative models, has seen remarkable advances with the introduction of models like Sora[1], Lumiere[2], StableVideoDiffusion[3] and CogVideoX[4]. These models show great promise in producing high-quality, longer-duration videos that adhere to physical laws.

In T2V research, the instruction-following capability is a key metric, reflecting the consistency between the generated video and its input text. Despite ongoing improvements in video clarity, aligning generated content with textual descriptions remains challenging. Inconsistencies, or even hallucinations, between the video and text can greatly undermine the quality of the output.

Therefore, accurately assessing the instruction-following performance of video generation models is crucial. Given the high cost and limited scalability of manual evaluations, we propose an end-to-end reward model to automate the evaluation of this capability in text-generated videos.

2 Related Work

Significant advancements have been made in image understanding and scoring, with models like CLIP[5] and ImageReward[6] contributing notably to image comprehension and evaluation.

However, the temporal dimension in videos, along with substantial changes across frames, complicates the representation of video-text alignment compared to images. While video understanding models, such as dual-stream models[7][8], I3D[9] and CogVLM2-video[10], have made considerable progress in video comprehension and keyframe extraction, studies indicate that general understanding models often fail to align with human preferences in video scoring tasks and struggle with effective automatic alignment scoring (e.g.,TIGER-Lab/GenAI). This discrepancy may arise from the absence of a rigorously defined scoring standard, which requires a well-annotated dataset to guide model scoring. Additionally, the features learned by multimodal understanding models are typically expressed as vocabulary, essentially treating scoring as an ongoing classification task, which may not be optimal in this context.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

To mitigate this issue, some researchers have attempted supervised learning fine-tuning on well-annotated datasets; however, the evaluation of alignment indicators (e.g., VideoScore–GenAI Tiger-Lab) remains inaccurate.

3 Study Proposal

This paper aims to develop a reward model that quantifies the alignment between images, videos, and textual descriptions, facilitating automated scoring.

3.1 Datasets Preparation

We will use VidProM[11], a dataset containing extensive T2V pairs from different models.

3.2 Base Model

We plan to utilize the latest image and video understanding models, leveraging their robust comprehension capabilities, and perform supervised training on annotated video-text pairs datasets to derive a score that reflects the alignment degree of text-image pairs.

3.3 Training Method

We plan to utilize the latest image and video understanding models, leveraging their robust comprehension capabilities, and perform supervised training on annotated video-text pairs datasets to derive a score that reflects the alignment degree of text-image pairs.

While current image and video understanding models are not yet fully mature and may exhibit hallucination issues, our task requires only the generation of a numerical value representing alignment, and we believe these models can effectively accomplish this scoring task.

The proposed reward model not only advances the automated evaluation of multimodal alignment but also establishes a foundation for the future application of reinforcement learning methods, such as PPO and DPO, to improve the trajectory of AI-generated videos.

References

[1] OpenAI. *Video Generation Models as World Simulators*. Available at: `https://openai.com/index/video-generation-models-as-world-simulators/`, 2024. Accessed: October 20, 2024.

[2] O. Bar-Tal, H. Chefer, O. Tov, et al. *Lumiere: A space-time diffusion model for video generation*. arXiv preprint arXiv:2401.12945, 2024.

[3] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, et al. *Stable video diffusion: Scaling latent video diffusion models to large datasets*. arXiv preprint arXiv:2311.15127, 2023.

[4] Z. Yang, J. Teng, W. Zheng, et al. *Cogvideox: Text-to-video diffusion models with an expert transformer*. arXiv preprint arXiv:2408.06072, 2024.

[5] A. Radford, J. W. Kim, C. Hallacy, et al. *Learning transferable visual models from natural language supervision*. In International Conference on Machine Learning, pages 8748–8763. PMLR, 2021.

[6] J. Xu, X. Liu, Y. Wu, et al. *Imagereward: Learning and evaluating human preferences for text-to-image generation*. Advances in Neural Information Processing Systems, 2024, 36.

[7] E. Fish, J. Weinbren, A. Gilbert. *Two-stream transformer architecture for long video understanding*. arXiv preprint arXiv:2208.01753, 2022.

[8] R. Liu, Y. Fang, F. Yu, et al. *Deep video understanding with video-language model*. In Proceedings of the 31st ACM International Conference on Multimedia, pages 9551–9555, 2023.

[9] X. Wang, Y. Zhang, Y. Wang, et al. *Quo vadis, action recognition? A new model and the Kinetics dataset*. arXiv preprint arXiv:2301.12345, 2023.

[10] W. Hong, W. Wang, M. Ding, et al. *Cogvlm2: Visual language models for image and video understanding*. arXiv preprint arXiv:2408.16500, 2024.

[11] W. Wang, Y. Yang. *VidProm: A million-scale real prompt-gallery dataset for text-to-video diffusion models*. arXiv preprint arXiv:2403.06098, 2024.