Is Your World Simulator a Good Story Presenter? A Consecutive Events-Based Benchmark for Future Long Video Generation

Supplementary Material

1. Related Works

1.1. Text-to-Video Generative Models

Recent advancements in vision generation have been largely dominated by diffusion models [4, 14, 22, 29, 37, 39, 41]. The open-source release of Stable Diffusion [24] has catalyzed extensive research and development in this field, while closed-source models have achieved commercial-grade quality [7, 12, 23, 25]. In parallel, another promising research direction involves training autoregressive models for video generation, where models are trained to sequentially predict tokens for video generation [1, 5, 11, 13, 19, 31–34, 38]. Despite these advancements, challenges remain in the long-term event continuity and logical narrative progression. Some contemporaneous work [36] suggests certain method to alleviate this significant problem, but there is still no standard way to quantify models' capacity in story presentation, and our creative-style prompts can still be challenging for them. Our benchmark provides a testbed for these lines of research by emphasizing the generation of coherent, long videos grounded in consecutive events.

1.2. Video Generation Metrics and Benchmarks

The rapid advancements in text-to-video (T2V) generation have driven the development of various benchmarks to evaluate the performance of generative models across diverse dimensions [6, 9, 10, 17, 18, 20, 40]. Traditional metrics like FVD [28] and IS [26] have been widely used to assess frame quality and text-frame alignment but fall short in capturing critical aspects such as subject consistency, temporal coherence, and physical commonsense correctness, particularly for novel or dynamic scenes [2]. Recently, benchmarks like VBench [9] and EvalCrafter [17] have expanded the evaluation scope, incorporating advanced metrics for dynamic attributes and human ratings. Specialized benchmarks such as T2V-CompBench [27] and DEVIL [16], PhyGenBench [21] address compositional and dynamic characteristics and physical realism. However, previous methods largely focus on evaluating detailed dynamics within short videos, failing to address the broader challenge of generating coherent, story-driven long videos. To bridge this gap, our method introduces a holistic evaluation framework that emphasizes narrative completion and the accurate depiction of sequential events, providing a robust benchmark for assessing long video generative models. We note that there are also some story-related video generation works [3, 8]. However, they focus more on enhancing control over the generation process or improving story visualization, which differs from our approach.

2. Supplementary Part of StoryEval

2.1. More Details about Prompt Suite

Retrieval-Based Prompts. Here we claim that the retrieval-based prompts ("Retrieval" class) we select are designed to match the general condition as illustrated in main paper. The lengths of videos retrieved from real-world are always hard to control, but the major part of the retrieval-based videos we select cover 5 to 20 second length, and for the videos with longer time span (about 1 minute), their events are also simple and direct enough to be completed in 10 seconds. Some of the retrieved videos are slow shots that can be accelerated.

Examples that GPT-4o rejects to answer. We note that for some prompts, GPT-4o always reject due to the security concern about "self-harm", "violence", and "sexual" issues in the generated videos, and we filter the prompts that may induce evaluation models to obtain unsafe videos. Nevertheless, we claim that the prompts themselves are manually checked and safe, and for all the prompts before post-processing filtration, there are always some closed-source models be able to generate safe videos. We think the reason for the security concerns may be that for most of the open-source model, their generated videos are still not that smooth and accurate, sometimes resulting unclear items and actions, and there may be misaligned movements that considered unsafe by GPT-4o. Examples of some prompts that are filtered because of this are shown in Figure 1. (Content Warning: the videos in Figure 1 are judged by GPT-4o to be potentially uncomfortable (associated with self-harm, violence, or sexual content, although human may disagree with that), please watch with caution.)

Note that as illustrated in post-processing filtration part in the main paper, we just filter the prompts which let at least two models (except ModelScope) fail to get the completion rates, there still exist some prompts which let the models fail to generate videos or make GPT-4o rejects to answer. In this case, we just let the completion rates on these cases to be zero and calculate the **Non-Response Rate** for each model, which evaluates how probable it is that the video model will fail to get a result for GPT-4o out of 423 prompts. The full experimental results containing Non-Response Rate are in Table 3.

2.2. More Details about Evaluation Process

Discussion of VLM verifiers. In this paper, we choose LLaVA-OV-Chat-72B rather than its lightweight version LLaVA-OV-Chat-7B, to evaluate the video generative models. The reason is that facing the videos generated by open-

Benchmark	Type	Evaluate Closed Source Models	Prompt Style		Evaluation	
			Retrieval	Creative	Longer Videos	Story Evaluation
FETV [18]	General	✗	✓	✓	✗	✗
EvalCrafter [17]	General	✓	✓	✓	✗	✗
T2VBench [10]	Temporal Dynamics	✓	✗	✗	✗	✗
TC-Bench [6]	Temporal Composition	✗	✓	✓	✗	✗
Chronomagic [40]	Time-Lapse	✓	✓	✗	✓	✗
T2V-Comp [27]	General Composition	✓	✗	✓	✓	✗
VBench [9]	General	✓	✗	✓	✓	✗
StoryEval (Ours)	Consecutive Events	✓	✓	✓	✓	✓

Table 1. **Comparisons to existing T2V benchmarks.** No previous works consider story evaluation. "Longer Videos" means the benchmark evaluate some videos with at least 5-seconds length.

source models, which may be noisier than the real-world videos or those generated by closed-source models, a larger 72B model performs much better than the 7B model, even if they always have similar results on the high quality videos. We can view the low-quality generated videos as harder visual understanding task, so the robustness of VLM verifiers is critical for all the VLM-based video generation benchmarks [9, 17, 20, 27]. Therefore, we choose the strongest model as verifier to ensure the accuracy of our StoryEval benchmark.

Query Prompts. In our StoryEval benchmark, we use two-step querying to obtain the completion rates. Here we use blue words to denote the places that differ for each prompt and video, and red to show the processed key frames that are sent to VLM verifier together with the second query texts.

Step1: Describe the video clips

Please describe the given key frames in the video in detail, in temporal order. The video may be generated by some video generative model rather than sampling from the real world, so it may be vague or not clear. You can point out if you don't see the video clearly.

Step2: Get completion rates

This is the description of the video you generated before, please refer to it to complete the following tasks.
{ The key frames from the video depict a sequence involving a bear near a waterfall, moving from a rock into the water. Here is the detailed temporal order of the frames:
1. The bear is standing on a rock near the edge of... }

Now, based on these descriptions and the video, you are asked to accurately determine if the following generated video fulfills the requirements of the prompt. The prompt contains several (2∼4) events, you need to judge if each event is strictly completed in the video. If the event is completed, please mark it as 1, otherwise, mark it as 0.

For example, if the prompt is: "A man dribbles a basketball and then throws it in a court", the prompt describes two events: "A man dribbles a basketball" and "And then the man throws the basketball in a court". But if the video generated using this prompt only accomplishes dribbling or only accomplishes shooting, then the completion list is [1, 0]. If you think both events are not completed, the completion list is [0, 0], etc.

Please judge whether the event are completed very strictly. If you think an item is blurry, hard to identify, or the action is vague, you should judge it as not completed. And please explain the reasons in detail before you give out the score.
You also need to check the item consistency between different events. If the prompt implies that the subject (or object) in different events should be the same, but in the video they are different, you should mark the later event in the prompt as not completed. For example, for the above prompt, if the man that dribbles the basketball is different from the man that throws the basketball (should be the same people, but video shows two different people), or the basketball that's dribbled is different from the basketball that's thrown (should be the same object, but video shows two different objects), you must mark the later event 'throwing the ball' as not completed.
Remember, you should judge whether the events are completed very strictly. And you should first provide the reasons or analysis for each event, and then give out the list of completion flag for each event (0 or 1 for uncompleted or completed).
Please remember to output the complete list at the end of output again, strictly follow the format: 'Finally we have [COMPLETE_LIST]: 1, 0' in a single line.

Now, let's begin scoring! The prompt is '{A bear walks by a waterfall, slips its foot, and then falls off a cliff.}', there are {3} events:
{ 1. A bear walks by a waterfall
2. The bear slips its foot
3. And then the bear falls off a cliff }
[Processed frames from generated videos appended]

2.3. Comparison with Previous Benchmarks

We compare StoryEval with previous T2V benchmarks and show the results in Table 1.

2.4. Discussion about Prompt Size

We claim that the size of StoryEval (423) is enough for stable evaluation, as current works (e.g. FETV (619), VBench (100 per metric), EvalCrafter (700) TC-Bench (120)) use similar amounts of prompts.

2.5. Discussion about Video Length

Although the length of generated videos will be longer in the future, we think this will not affect our evaluation too much. LongVideoBench [35] shows top VLMs (e.g. GPT-4o/LLaVA-OV-72B) perform well on both 8s-15s and 15s-60s (low-level) video tasks, which is longer than current generated videos (10s). Besides, StoryEval focuses on high-level events, which is easier to judge and has lower requirements on temporal resolution. So we can relatively reduce frame sampling rates for longer videos.

2.6. Discussion about Post Filtration

We mention that we filter the data which are rejected by GPT-4o mainly to ensure the stability of the evaluation. The number of prompts for this is less than 20, and we validate that including them has very little influence on the final performance.

3. Supplementary Part of Experiments

3.1. Model details about experimental settings.

Video Length. In Table 2, we show the length of generated videos of each model in our experiments. Here for open-source models, we increase the length of the video generation to 10 seconds as much as possible while keeping the generation stable. For example, for Pyramid-Flow, if the generation length is longer than 7 seconds, some examples may become blurred at the end, so we set the number of frames such that the generation length is about 7 seconds.

Number of key frames. Note that different generative models have quite different number of key frames K in each video, we use the following formula to select the number of key frames used for evaluating the videos.

$$\text{num of frames} = \max(\min(32, \lfloor K/4 \rfloor), 4) \qquad (1)$$

Since our prompts consider short stories including simple events, as well as the generative models generally generates videos that present slow motion, this sampling strategy can be sufficient for accurate evaluation results while keeping reasonable cost for GPT-4o querying.

Length of prompts. In detail, the average number of the characters in the prompts in StoryEval is 83.4. Note that we mainly consider the story or events of prompt, rather than the detailed description of states, which always contain the adjectival details like color. So the length of prompts are always quite simple (as shown in the introduction figures in the main paper). Besides, some models like CogVideoX may over-fit to the long detailed prompts for generating good videos, and will perform worse on short prompts we provide. However, for fair comparison, we just use the same prompts without LLM refining in StoryEval, and expect to use long prompts suite in the next StoryEval version.

Model	FPS	Num. of Frames	Length (s)
Pika-1.5	24.0	120	5.0
Hailuo	25.0	141	5.6
Kling-1.5	30.0	313	10.4
ModelScope	8.0	16	2.0
EasyAnimate v4	14.0	144	10.3
Open-Sora-Plan 1.3.0	18.0	93	5.2
Open-Sora 1.2	24.0	255	10.6
VideoCrafter2	10.0	100	10.0
CogVideoX-5B	8.0	49	6.1
Vchitect-2.0	8.0	80	10.0
Pyramid-Flow	24.0	169	7.0

Table 2. **Statistics of generated videos from different models.** We can see that longer generation length (second) or more number of frames don't necessarily ensure better StoryEval performance.

3.2. Detailed Results

In this section, we show additional experimental results.

In Table 3, we show the Non-Response Rate defined in Sec. 2.1 for GPT-4o verifier.

In Table 4, we append the average result of VBench, which still shows ranking difference with StoryEval in Pyramid-Flow.

In Table 5, we further extend Table 4 and show the performance of different models on different benchmarks.

Finally in Table 6, we see that on almost all the sub-metrics, LLaVA-OV-Chat-72B has similar ranking results to GPT-4o.

References

[1] Meta AI. Movie-gen. https://ai.meta.com/research/movie-gen/, 2024. Accessed: 2024-11-09. 1

[2] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems*, 35:31769–31781, 2022. 1

[3] Emanuele Bugliarello, H Hernan Moraldo, Ruben Villegas, Mohammad Babaeizadeh, Mohammad Taghi Saffar, Han Zhang, Dumitru Erhan, Vittorio Ferrari, Pieter-Jan Kindermans, and Paul Voigtlaender. Storybench: A multifaceted benchmark for continuous story visualization. *Advances in Neural Information Processing Systems*, 36:78095–78125, 2023. 1

Open-Source Model	Human	Animal	Object	Retrieval	Creative	Easy	Hard	Non-R. Rate ↓	Average
Pika-1.5 [23]	17.1%	17.9%	22.1%	26.9%	20.6%	40.3%	4.4%	**0.2%**	19.4%
Hailuo [7]	**38.2%**	38.3%	27.5%	**42.6%**	18.0%	58.9%	9.7%	0.9%	35.1%
Kling-1.5 [12]	37.2%	**44.9%**	**36.6%**	39.4%	**36.0%**	**60.8%**	**16.4%**	1.7%	**40.1%**
Closed-Source Model									
ModelScope [30]	11.5%	8.8%	8.2%	9.9%	5.2%	21.7%	5.0%	7.8%	9.8%
EasyAnimate v4 [37]	15.6%	12.8%	11.4%	18.4%	8.2%	26.7%	3.6%	1.7%	13.3%
Open-Sora-Plan 1.3.0 [14]	9.1%	9.7%	9.4%	13.2%	7.1%	18.2%	3.2%	0.2%	9.4%
Open-Sora 1.2 [41]	16.4%	18.3%	16.2%	**24.7%**	11.8%	32.7%	4.3%	**0.0%**	17.9%
VideoCrafter2 [4]	11.0%	13.1%	9.8%	18.0%	5.5%	29.9%	1.7%	0.9%	12.2%
CogVideoX-5B [39]	17.1%	16.4%	14.0%	16.0%	7.4%	35.4%	**4.6%**	5.2%	16.4%
Vchitect-2.0 [29]	**21.5%**	**19.9%**	**20.4%**	22.0%	**15.2%**	**42.8%**	3.9%	1.7%	**21.7%**
Pyramid-Flow [11]	17.8%	16.5%	12.8%	23.4%	9.7%	35.1%	1.0%	0.5%	16.0%

Table 3. **Full StoryEval evaluation results on 11 video generative models with GPT-4o verifier.** Non-Response Rate ("Non-R. Rate") assesses the likelihood that the video model will fail to obtain completion rates from GPT-4o in 423 prompts.

Model	VBench Temp. Cons.	VBench Avg.	StoryEval Avg.
Kling1.5	$\geq$ 97.93% (1)	$\geq$ 81.85% (1)	40.1% (1)
Pika1.5	$\geq$ 97.83% (2)	$\geq$ 80.69% (4)	19.4% (3)
ModelScope	90.88% (8)	75.75% (8)	9.8% (7)
Open-Sora 1.2	96.85% (5)	79.23% (6)	18.2% (4)
Open-Sora-Plan 1.3.0	97.28% (4)	77.23% (7)	9.4% (8)
VideoCrafter2	95.67% (6)	80.44% (5)	12.2% (6)
Vchitect-2.0	95.03% (7)	81.57% (3)	21.7% (2)
Pyramid-Flow	97.56% (3)	81.72% (2)	16.0% (5)

Table 4. **StoryEval can be an effective complementary metric to the traditional detail-oriented metrics.** Here "VBench Avg." means the average value of VBench over and "Temp. Cons." denotes VBench's temporal consistency submetric, which including the evaluation of subject consistency, background consistency, temporal flickering and motion smoothness.

[4] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. 1, 4

[5] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021. 1

[6] Weixi Feng, Jiachen Li, Michael Saxon, Tsu-jui Fu, Wenhu Chen, and William Yang Wang. Tc-bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation. *arXiv preprint arXiv:2406.08656*, 2024. 1, 2

[7] HailuoAI. Hailuo. https://hailuoai.video/, 2024. Accessed: 2024-11-09. 1, 4

[8] Panwen Hu, Jin Jiang, Jianqi Chen, Mingfei Han, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. Storyagent: Customized storytelling video generation via multi-agent collaboration. *arXiv preprint arXiv:2411.04925*, 2024. 1

[9] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 1, 2

[10] Pengliang Ji, Chuyang Xiao, Huilin Tai, and Mingxiao Huo. T2vbench: Benchmarking temporal dynamics for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5325–5335, 2024. 1, 2

[11] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. 2024. 1, 4

[12] KlingAI. Kling. https://klingai.kuaishou.com/, 2024. Accessed: 2024-11-09. 1, 4

[13] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 1

[14] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, 2024. 1, 4

[15] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and

Result (rank)	VBench		T2V-Comp		Chronomagic-Bench		StoryEval
	Temp. C.	Avg.	Dynamic	Avg.	CHScore	GPT4o-MTScore	Avg.
CogVideoX-5B	94.4 (4)	81.6 (1)	27.4 (1)	41.9 (1)	48.5 (4)	3.36 (1)	16.4 (2)
ModelScope	90.9 (5)	75.8 (5)	20.7 (5)	31.5 (5)	-	-	9.8 (4)
Open-Sora-Plan 1.3.0	97.3 (1)	77.2 (4)	23.4 (4)	36.7 (4)	71.0 (2)	2.64 (3)	9.4 (5)
Open-Sora 1.2	96.9 (2)	79.8 (3)	25.0 (2)	38.5 (3)	51.6 (3)	2.56 (4)	17.9 (1)
VideoCrafter2	95.7 (3)	80.4 (2)	24.6 (3)	40.2 (2)	80.1 (1)	2.68 (2)	12.2 (3)

Table 5. **Different T2V models on different benchmarks.** "Temp. C.": Temporal Consistency. CHScore: evaluates temporal coherence. GPT4o-MTScore: evaluates metamorphic attributes.

Model	Human	Animal	Object	Retrieval	Creative	Easy	Hard	Average
Pika-1.5	23.9	23.8	26.5	35.2	22.9	41.1	13.0	25.0
Hailuo	**48.0**	40.1	**35.6**	**51.7**	19.5	**58.3**	17.1	41.0
Kling-1.5	41.9	**46.0**	35.1	41.7	**30.8**	56.1	**24.1**	**41.7**
ModelScope	17.1	13.1	13.2	13.7	6.2	25.3	4.9	15.2
EasyAnimate v4	21.7	18.1	15.5	21.2	10.5	29.6	5.0	18.5
Open-Sora-Plan 1.3.0	13.5	13.2	9.6	17.1	6.9	28.3	2.2	12.7
Open-Sora 1.2	26.2	22.2	20.2	32.2	15.4	37.8	10.8	23.6
VideoCrafter2	15.2	14.2	12.8	15.3	8.0	33.0	4.7	14.7
CogVideoX-5B	19.7	20.7	17.4	27.2	8.1	37.6	7.1	19.9
Vchitect-2.0	**33.4**	**30.5**	**33.6**	**33.6**	**20.5**	**51.4**	**19.1**	**31.6**
Pyramid-Flow	23.6	20.0	15.8	26.4	10.5	38.1	4.5	20.3
Spearman's ρ	0.91	0.97	0.96	0.89	0.98	0.96	0.71	0.98
Kendall's τ	0.81	0.89	0.86	0.78	0.92	0.86	0.53	0.93

Table 6. **Full StoryEval evaluation results on 11 video generative models with LLaVA-OV-Chat-72B [15] verifier.** The ρ and τ are evaluated between the completion rates of GPT-4o and that of LLaVA-OV-Chat-72B model. We can see that on almost all the submetrics, LLaVA-OV-Chat-72B has similar ranking results to GPT-4o. Here LLaVA-OV-Chat-72B will response all the generated videos, so we skip Non-Response Rate.

Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 5

[16] Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective. *arXiv preprint arXiv:2407.01094*, 2024. 1

[17] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 1, 2

[18] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[19] Gary Marcus, Ernest Davis, and Scott Aaronson. A very preliminary analysis of dall-e 2. *arXiv preprint arXiv:2204.13807*, 2022. 1

[20] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024. 1, 2

[21] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024. 1

[22] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1

[23] Pika. Pika. https://pika.art/, 2024. Accessed: 2024-11-09. 1, 4

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[25] RunwayML. Gen-3. https://app.runwayml.com/, 2024. Accessed: 2024-11-09. 1

[26] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques

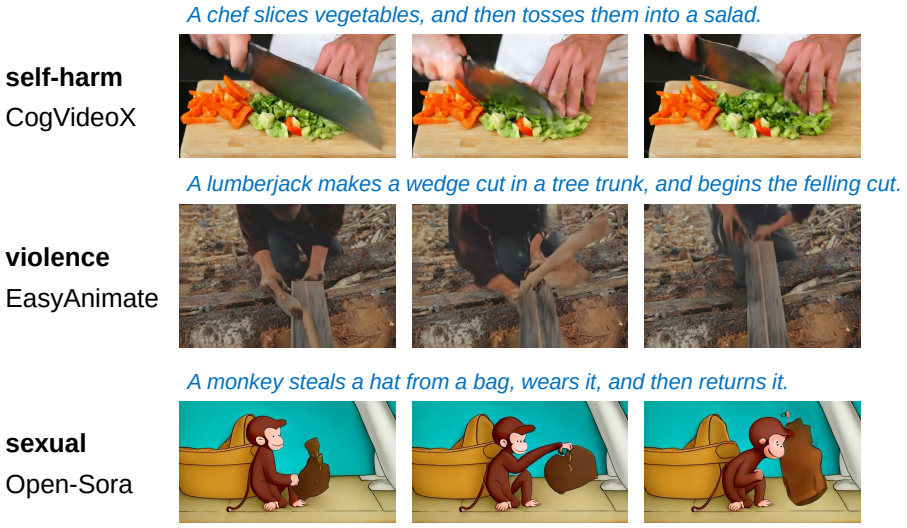

Figure 1. **Some prompts that have their corresponding videos rejected by GPT-4o due to security concerns.** Samples are selected from three open-source models. Content Warning: These videos are judged by GPT-4o to be potentially uncomfortable (associated with self-harm, violence, or sexual content, although human may disagree with that), please watch with caution.

for training gans. *Advances in neural information processing systems*, 29, 2016. 1

[27] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. *arXiv preprint arXiv:2407.14505*, 2024. 1, 2

[28] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 1

[29] Vchitect. Vchitect 2.0. https://github.com/Vchitect/Vchitect-2.0, 2024. Version 2.0, Accessed: 2024-11-09. 1, 4

[30] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 4

[31] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 1

[32] Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. Loong: Generating minute-level long videos with autoregressive language models. *arXiv preprint arXiv:2410.02757*, 2024.

[33] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*, 2019.

[34] Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, Chong Luo, Yueyi Zhang, and Zhiwei Xiong. Art-v: Auto-regressive text-to-video generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7395–7405, 2024. 1

[35] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024. 3

[36] Ziyi Wu, Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Yuwei Fang, Varnith Chordia, Igor Gilitschenski, and Sergey Tulyakov. Mind the time: Temporally-controlled multi-event video generation. *arXiv preprint arXiv:2412.05263*, 2024. 1

[37] Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. Easyanimate: A high-performance long video generation method based on transformer architecture. *arXiv preprint arXiv:2405.18991*, 2024. 1, 4

[38] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 1

[39] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 4

[40] Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *arXiv preprint arXiv:2406.18522*, 2024. 1, 2

[41] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 1, 4