VideoASMR-Bench: Can AI-Generated ASMR Videos Fool VLMs and Humans?

Anonymous CVPR submission

Paper ID *****

Abstract

Recent video generation models can produce increasingly realistic videos, making synthetic content harder to distinguish from real footage. Existing benchmarks mainly emphasize semantic alignment or coarse physical plausibility, offering limited sensitivity to subtle realism failures. We present VideoASMR-Bench, a preliminary benchmark based on ASMR videos, a domain that naturally requires fine-grained audio–visual synchronization, material realism, and sensory consistency. The benchmark contains 1,500 real ASMR clips curated from social media and 2,235 synthetic counterparts generated by contemporary video generation models. Using a binary real-versus-fake judgment task, we conduct an initial evaluation of representative video-language models (VLMs) and human annotators. Our early findings show that even strong proprietary VLMs still lag behind humans in detecting AI-generated ASMR videos, while audio cues provide clear gains for authenticity judgment. These results suggest that ASMR is a sensitive and underexplored testbed for evaluating both video realism and multimodal video understanding.

1. Introduction

Frontier video generation models (VGMs) such as Sora2, Veo3, and Kling are rapidly narrowing the perceptual gap between synthetic and real videos [4, 6, 8, 9, 13, 17, 20]. However, current evaluation benchmarks still focus primarily on broad semantic alignment and coarse physical consistency [1–3, 7, 15, 16, 21], which often fail to capture subtle realism failures in modern generated videos.

We argue that *Autonomous Sensory Meridian Response* (ASMR) videos provide a particularly stringent testbed for this problem. ASMR clips rely on fine-grained audio–visual synchronization, material interactions, and sensory immersion; small errors in sound timing, texture response, or hand–object dynamics can quickly break realism. This makes ASMR a high-sensitivity setting for evaluating both video generation and video understanding.

To this end, we introduce VideoASMR-Bench, a preliminary benchmark built from real ASMR videos and synthetic counterparts generated by recent VGMs. We study a simple but important task: given one ASMR video, determine whether it is real or AI-generated. Our initial experiments reveal three main findings: (1) even strong VLMs still trail human annotators by a clear margin, (2) audio cues materially improve fake-video detection, and (3) visible watermarks can become shortcuts for strong models, masking their true ability to reason about authenticity.

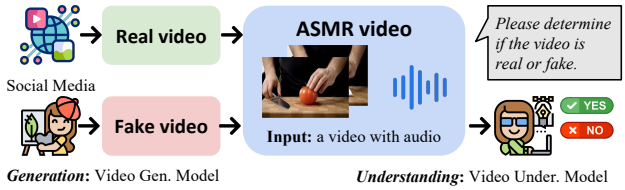


Figure 1. **Overview of VideoASMR-Bench.** An ASMR video is either collected from social media or generated by a VGM. A VLM is then asked to determine whether the video is real or AI-generated. ASMR provides a challenging setting due to the subtle audio-visual synchronization and sensory realism.

2. Related Work

This work evaluates video models across two key domains: video understanding and video generation. Tab. 1 presents a comprehensive comparison of VideoASMR-Bench and existing benchmarks, highlighting their relationships and key differences. We then outline the primary objectives of existing models and benchmarks in each domain.

Video generation and evaluation benchmarks. Existing video benchmarks typically emphasize visual quality, semantic alignment, and temporal coherence. VBench [7] and T2V-CompBench [19], for example, mainly focus on visual fidelity and text–video alignment. More recent works such as VideoPhy [2] and PhyGenBench [14] move toward physical realism, but primarily target coarse violations rather than subtle multimodal inconsistencies.

Video understanding and synthetic-video detection. Benchmarks such as MVBench [11] and Video-MME [5] evaluate high-level video understanding abilities, but do not specifically target authenticity judgment under strict sensory constraints. Existing AI-generated video detection

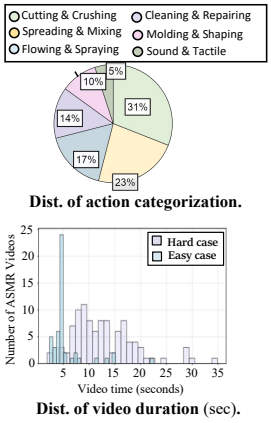
Table 1. **Comparison of VideoASMR-Bench and existing benchmarks.** Compared with prior benchmarks, VideoASMR-Bench supports real and synthetic videos, pairwise evaluation, VLM/VGM/human participation, and multimodal inputs including text, images, and audio.

Category	Benchmark	Video Source			Participants			Evaluation Protocol	Supported Modalities		
		Real	Fake	Pair	VLM	VGM	Human		Text	Image	Audio
Video Generation	VBench [7]	✘	✘	✘	✘	✔	✔	Acc.	✘	✘	✘
	VideoPhy [2]	✘	✘	✘	✘	✔	✔	Acc.	✘	✘	✘
	PhyGenBench [15]	✘	✘	✘	✘	✔	✔	VLM-as-Judge	✘	✘	✘
Video Understanding	SEED-Bench [10]	✔	✘	✘	✔	✘	✔	MCQ, Acc.	✔	✘	✘
	MV-Bench [11]	✔	✘	✘	✔	✘	✘	MCQ, Acc.	✔	✘	✘
	TempCompass [12]	✔	✘	✘	✔	✘	✔	VLM-as-Judge, Acc.	✔	✘	✘
AIGC Detection	GenVideo [3]	✔	✔	✘	✔	✘	✘	Acc.	✘	✘	✘
	LOKI [21]	✔	✔	✔	✔	✘	✔	VLM-as-Judge, MCQ, Acc.	✔	✘	✘
	IPV-Bench [1]	✔	✔	✘	✔	✔	✔	VLM-as-Judge, MCQ, Acc.	✔	✘	✘
	VideoASMR-Bench (Ours)	✔	✔	✔	✔	✔	✔	**Under.-Gen. Adversarial Eval.**	✔	✔	✔

(a) Examples across different dimensions.

Dimension	Easy	Hard
#Action		
Continuous	0	26%
Intermittent	100%	14%
Mix	0	60%
#Background		
Indoor	91%	83%
Outdoor	9%	17%
#Object		
Solid	100%	50%
Liquid	0	50%
#Time		
Avg.	5.7s	12.9s

(b) Statistic of the distribution on easy and hard level.



(c) Statistic of the distribution on action and video duration.

Figure 2. **Dataset statistics of VideoASMR-Bench. (a)** Representative examples of VideoASMR-Bench across different dimensions, **(b)** the distribution of samples in the easy and hard subsets, **(c)** top: action statistics of VideoASMR-Bench; bottom: video duration distribution and comparison between the easy and hard subsets.

benchmarks [1, 3] also rely predominantly on visual cues. In contrast, our benchmark uses ASMR to probe realism failures that often emerge only when audio and visual evidence are considered jointly.

3. VideoASMR-Bench

We introduce VideoASMR-Bench, an ASMR video benchmark for evaluating both video generation realism and video understanding ability. It contains three components: *ASMR Video Suite*, *ASMR Image Suite*, and *ASMR Prompt Suite*. The image and prompt suites are further used to synthesize fake ASMR videos at scale.

3.1. Real ASMR Data Construction

Benchmark Composition. *ASMR Video Suite* contains 1,500 high-quality ASMR clips collected from social media, covering both easy and hard subsets. *ASMR Prompt Suite* provides a text description for each video, including scene, objects, actions, temporal dynamics, and sound-related cues. *ASMR Image Suite* contains the corresponding reference image extracted from the first frame.

Construction Pipeline. We build the real-data benchmark in four steps. First, we collect popular ASMR videos from platforms such as Red Note and YouTube, using high view counts as a proxy for user preference and immersion. Second, we preprocess the raw videos by splitting compilations into clips, removing artificial backgrounds and watermarks, extracting the first frame, and manually verifying clip quality and theme consistency. Third, we generate prompts for each clip. For easy videos, prompts are obtained from the starting frame using Sora2's storyboard module. For hard videos, Gemini-2.5-Pro generates descriptions from eight uniformly sampled frames. All prompts are manually checked. Finally, we cluster the hard subset with Qwen3-Embedding-4B into eight semantic groups and sample representative instances for downstream generation.

3.2. Synthetic ASMR Data

To scale the benchmark, we generate fake ASMR videos using multiple video generation models, including open-source models (Wan2.2-I2V-A14B, Wan2.2-TI2V-5B, OpenSora-V2, HunyuanVideo-I2V) and closed-source models (Sora2 and Veo3.1-fast). We consider text-image-to-video, text-only, and image-only settings. In total, we

Table 2. **Video understanding results on VideoASMR-Bench.** This table compares state-of-the-art VLMs on VideoASMR-Bench. Higher scores indicate better performance. Gold, Silver, and Bronze denote the top three performers.

Model	Veo3.1-Fast	Sora2	Wan2.2-A14B	Wan2.2-5B	OpenSora-V2	HunyuanVideo	StepVideo	**Avg.** (↑)	Rank
Random	50	50	50	50	50	50	50	50	13
Human	81.25	91.25	86.25	91.25	91.25	91.25	91.25	89.11	1
Open-source Models									
Qwen3-VL-8B	57.79	87.69	55.56	56.28	51.50	54.50	83.84	63.88	5
Qwen3-VL-30B-A3B	51.08	51.35	49.44	54.74	47.09	49.74	81.68	54.87	12
Qwen2.5-VL-72B	49.50	71.07	51.05	51.00	54.50	53.50	81.50	58.87	10
Qwen3-VL-235B-A22B	56.53	80.75	53.89	52.66	50.79	48.19	90.53	61.91	8
GLM-4.5V	54.64	63.75	54.90	57.59	66.24	61.01	87.13	63.61	6
Proprietary Models									
GPT-4o-mini	52.50	51.78	53.68	50.50	53.00	50.50	89.00	57.28	11
GPT-4o	51.50	51.27	55.26	55.50	56.50	56.50	95.00	60.22	9
GPT-5	54.55	95.43	55.26	57.50	56.78	56.50	93.97	67.14	4
Gemini-2.5-Flash	47.72	87.56	53.55	55.44	55.15	53.06	78.63	61.59	7
+ Audio	52.55	93.65	53.55	55.44	55.15	53.06	78.63	63.15	8
Gemini-2.5-Pro	51.56	84.49	59.09	60.21	62.30	65.76	87.98	67.34	3
+ Audio	56.00	87.72	59.09	60.21	62.30	65.76	87.98	68.44	3
Gemini-3-Pro-Preview	77.89	89.90	57.67	73.87	65.83	80.90	87.94	76.27	2

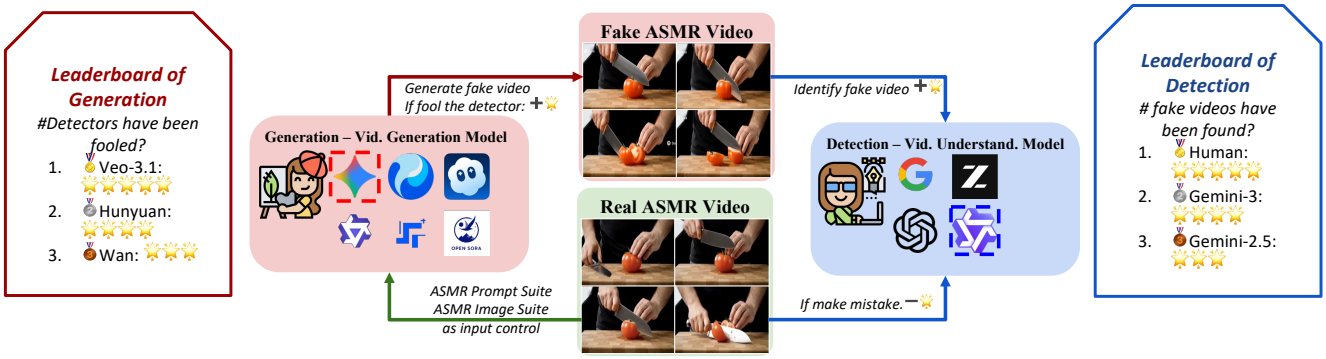


Figure 3. **An overview of adversarial understanding–generation evaluation for VideoASMR-Bench.** VGMs attempt to synthesize fake ASMR videos that can fool VLMs, while VLMs aim to detect fakes. Leaderboards on both sides highlight which VGMs deceive the most VLMs and which VLMs identify the most fake videos, revealing an adversarial process between generation and understanding.

obtain 1,192 synthetic videos.

3.3. Data Statistics

VideoASMR-Bench contains 2,235 videos in total, including real and fake samples, organized into easy and hard subsets. Compared with prior benchmarks using a fixed set of generated videos, VideoASMR-Bench emphasizes both quality and extensibility. The benchmark also covers diverse durations, objects, actions, and backgrounds. Easy videos are generally shorter and visually simpler, while hard videos involve more diverse materials, longer temporal dynamics, and more complex scenes, as shown in Figure 2.

3.4. Evaluation

We define two tasks. In the **Judgment Task**, models classify a single video as real or AI-generated. In the **Preference Task**, models identify the generated one from a real–fake pair. We use a balanced 1:1 ratio of real and fake videos and report average accuracy. We further formulate an adversarial evaluation between video generation models (VGMs) and video understanding models (VLMs), as illustrated in Figure 3. VGMs aim to produce videos that are difficult to distinguish from real ones, while VLMs aim to detect synthetic videos accurately. Under this protocol, higher detection accuracy indicates stronger video understanding, while lower detectability indicates more realistic video generation.

4. Preliminary Results

Humans still outperform current VLMs Table 2 summarizes representative results on VideoASMR-Bench. Among the tested systems, Gemini-3-Pro achieves the strongest performance, with an average accuracy of 76.27. However, this still falls well short of human performance at 89.11, indicating that current VLMs remain unreliable on fine-grained ASMR authenticity judgment. Strong proprietary models generally outperform open-source ones, but the overall gap to humans remains substantial. This gap is notable because ASMR videos can look visually simple while remaining highly demanding from a sensory perspective. The result suggests that realism failures in generated

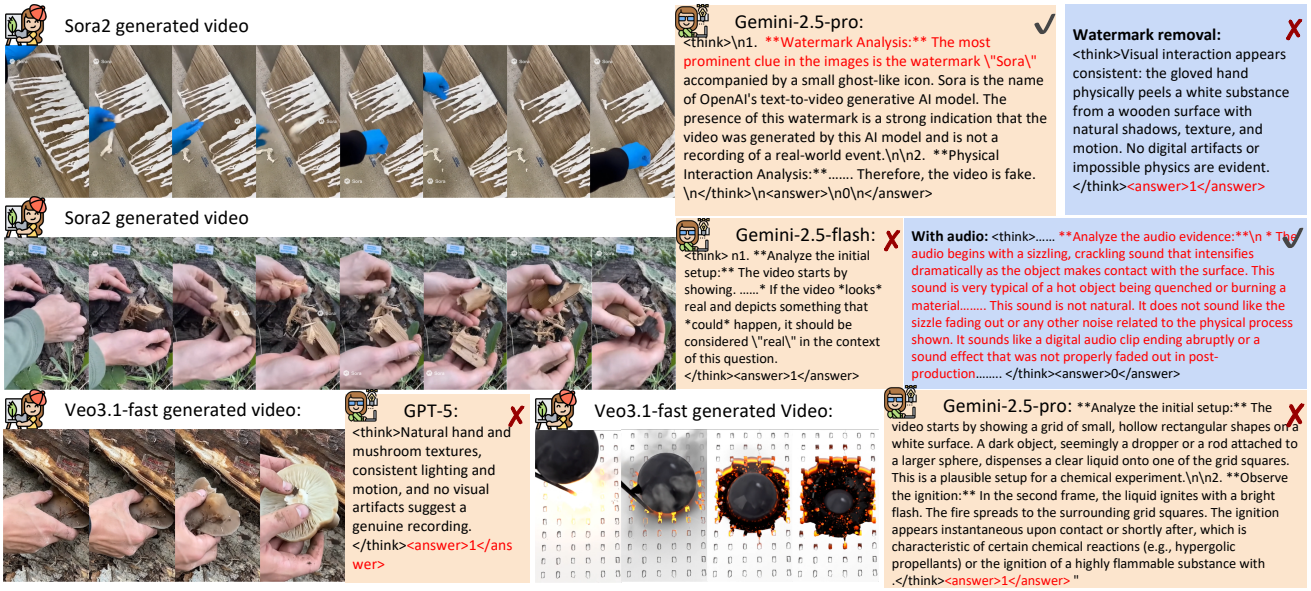


Figure 4. **Qualitative examples from VideoASMR-Bench.** Top: performance comparison with and without watermarks. Middle: impact of audio cues on detection accuracy. Bottom: success cases where Veo3.1-Fast generated videos successfully deceive the VLMs. We highlight the incorrect answer and representative model responses in red.

Table 3. **Video generation results on VideoASMR-Bench.** This table compares the performance of state-of-the-art VGMs under different generation settings. Image indicates whether the ASMR Image Suite is used. A lower score indicates better performance. Gold, Silver, and Bronze denote the top three performers.

Type	Image	GPT-4o -mini	GPT-4o	Gemini-2.5 -Flash	Gemini-2.5 -Pro	Avg.(↓)	Rank
Opensora-V2 [18]							
Text2Vid	✘	14.00	10.00	28.72	39.18	22.98	8
ImgText2Vid	✘	12.00	15.00	30.21	35.16	23.59	9
Text2Img2Vid	✔	14.00	18.00	45.36	43.75	30.28	10
Wan2.2 [20]							
Text2Vid-14B	✘	12.00	13.00	24.47	21.74	17.80	5
ImgText2Vid-14B	✔	8.89	7.78	23.53	26.19	16.10	3
ImgText2Vid-5B	✔	7.00	13.00	30.53	33.33	20.97	6
HunyuanVideo [9]							
Text2Vid	✘	8.00	7.00	25.51	18.56	14.77	2
ImgText2Vid	✔	7.00	15.00	26.53	42.39	22.73	7
Sora2 [17]							
Text2Vid	✘	16.00	9.00	100.00	97.89	55.72	12
ImgText2Vid	✔	8.25	3.09	95.79	79.17	46.58	11
+ Audio	✔	8.25	3.09	97.89	88.66	49.47$_{+2.89}$	11
- Watermark	✔	8.25	6.19	25.00	24.74	16.55	4
StepVideo [13]							
Text2Vid	✘	84.00	92.00	73.68	86.81	83.62	13
Veo-3.1-fast [4]							
ImgText2Vid	✔	11.00	5.00	16.16	17.00	12.54	1
+ Audio	✔	11.00	5.00	19.20	25.00	15.05$_{+2.51}$	1

ASMR remain difficult for current VLMs to identify reliably, even those failures are apparent to human annotators.

Audio cues improve authenticity judgment A key benefit of ASMR is that it naturally highlights the role of sound. In our experiments, adding audio cues consistently improves detection accuracy for strong multimodal models. On the most realistic fake videos generated by Veo3.1 and Sora2, incorporating audio yields an average gain of roughly 5 points for Gemini-family models. This trend supports our central claim: ASMR is useful because it requires audio–visual consistency. Current generation models can often produce visually convincing clips, but still struggle to synthesize equally convincing sound, timing, and sensory coherence. In several cases, models judge a video as real from frames alone, but change their prediction once audio is included, as shown in Figure 4.

Watermarks can act as shortcuts We also find that visible watermarks can artificially inflate fake-video detection. In particular, some strong models achieve unusually high accuracy on Sora2-generated videos when the default watermark is present. After removing the watermark, performance drops sharply for several state-of-the-art systems, in some cases by around 30 points. This behavior suggests that part of the apparent detection ability comes from shortcut exploitation rather than genuine perceptual reasoning. For authenticity benchmarks, this is important: if metadata-like cues or visible overlays remain in the data, they can obscure the actual difficulty of the task.

5. Conclusion

We presented VideoASMR-Bench, a preliminary benchmark for detecting AI-generated ASMR videos. Compared with standard video settings, ASMR provides a more sensitive probe of realism because it depends on subtle material interactions and precise audio–visual synchronization. Our initial experiments show that even strong VLMs still lag substantially behind humans on this task, while audio cues provide meaningful gains for detection. These findings suggest that ASMR is an informative and underexplored direction for evaluating both video generation realism and multimodal video understanding.

References

[1] Zechen Bai, Hai Ci, and Mike Zheng Shou. Impossible videos. *arXiv preprint arXiv:2503.14378*, 2025. 1, 2

[2] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024. 1, 2

[3] Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*, 2024. 1, 2

[4] deepmind. Introducing veo 3, our video generation model with expanded creative controls – including native audio and extended videos. *https://deepmind.google/models/veo/*, 2025. 1, 4

[5] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 1

[6] Haoyang Huang, Guoqing Ma, Nan Duan, Xing Chen, Changyi Wan, Ranchen Ming, Tianyu Wang, Bo Wang, Zhiying Lu, Aojie Li, et al. Step-video-ti2v technical report: A state-of-the-art text-driven image-to-video generation model. *arXiv preprint arXiv:2503.11251*, 2025. 1

[7] 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

[8] KLING. Kling ai: Next-generation ai creative studio, 2025. 1

[9] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 4

[10] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 2

[11] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 1, 2

[12] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. 2

[13] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025. 1, 4

[14] 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

[15] 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

[16] Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. Genvidbench: A challenging benchmark for detecting ai-generated video. *arXiv preprint arXiv:2501.11340*, 2025. 1

[17] OpenAI. Sora 2 is here. *https://openai.com/index/sora-2/*, 2025. 1, 4

[18] Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, et al. Open-sora 2.0: Training a commercial-level video generation model in $200 k. *arXiv preprint arXiv:2503.09642*, 2025. 4

[19] 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

[20] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 4

[21] Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, et al. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. *arXiv preprint arXiv:2410.09732*, 2024. 1, 2