T2AV-Compass: Towards Unified Evaluation for Text-to-Audio-Video Generation

Anonymous CVPR submission

Paper ID 2

Abstract

*Text-to-Audio-Video (T2AV) generation aims to synthesize temporally coherent video and semantically synchronized audio from natural language. However, its evaluation remains fragmented, often relying on unimodal metrics or narrow benchmarks that fail to capture cross-modal alignment, instruction following, and perceptual realism. To address this limitation, we present **T2AV-Compass**, a unified benchmark for comprehensive evaluation of T2AV systems. It consists of 500 diverse, complex prompts constructed via a taxonomy-driven pipeline to ensure semantic richness and physical plausibility. T2AV-Compass further introduces a dual-level evaluation framework that combines objective signal-level metrics with a subjective, MLLM-based protocol for instruction following and realism assessment. Extensive evaluation of 15 representative T2AV systems shows that even the strongest models still fall substantially short of human-level cross-modal consistency, with persistent failures in audio realism and fine-grained synchronization. These results position T2AV-Compass as a challenging diagnostic testbed for advancing multimodal generation.*

1. Introduction

Generative AI has witnessed a paradigm shift from unimodal synthesis to cohesive multimodal content creation [8, 11, 30, 31, 36], with Text-to-Audio-Video (T2AV) generation emerging as a frontier that unifies visual dynamics and auditory realism. While breakthroughs from proprietary systems like Sora [24] and Veo [6] to open efforts [17, 27, 41] demonstrate high-fidelity generation, **T2AV evaluation remains fundamentally underdeveloped.**

Existing benchmarks largely evolve from unimodal settings, prioritizing either isolated visual quality (e.g., VBench [12], EvalCrafter [20]) or audio fidelity (e.g., AudioCaps [14], AudioLDM-Eval [18]). Emerging joint evaluations often face critical trade-offs, including limited coverage of fine-grained coupling, insufficient handling of compositional prompts, and a lack of interpretable diagnostic signals. Consequently, current evaluations struggle to answer core questions: Do generated sounds precisely correspond to complex visual events? Do models faithfully follow detailed instructions while maintaining physical plausibility? High-quality T2AV generation requires simultaneous success in perceptual quality, cross-modal alignment, instruction following under compositional constraints, and commonsense realism, making evaluation substantially more challenging than in unimodal settings.

To address this gap, we introduce **T2AV-Compass** (Figure 1), the first comprehensive benchmark designed specifically for T2AV generation. Our contributions are threefold. **(1) Taxonomy-Driven High-Complexity Benchmark:** We curate 500 dense prompts via a hybrid pipeline of taxonomy-based design and real-world video inversion, targeting frequently overlooked fine-grained constraints such as off-screen sound and physical causality. **(2) Unified Dual-Level Evaluation Framework:** We integrate objective signal metrics (for video/audio quality and cross-modal alignment) with a checklist-based *MLLM-as-a-Judge* protocol (for instruction following and realism), enabling more interpretable and diagnostic evaluation. **(3) Extensive Benchmarking and Empirical Insights:** We systematically evaluate 15 representative T2AV systems. Our analysis reveals a clear "Audio Realism Bottleneck": current models still struggle to synthesize physically grounded audio textures that match their visual fidelity.

2. Data Curation and Evaluation

2.1. Data Construction

To ensure the diversity and complexity of our benchmark, we employ a three-stage construction pipeline comprising taxonomy-based prompt design, multi-source data collection, and real-world video inversion (Figure 2).

Data Collection. To establish broad semantic coverage, we aggregate prompts from diverse high-quality sources, including VidProM, the Kling AI community, LMArena, and Shot2Story [10, 15, 21, 37]. To mitigate long-tail imbal-

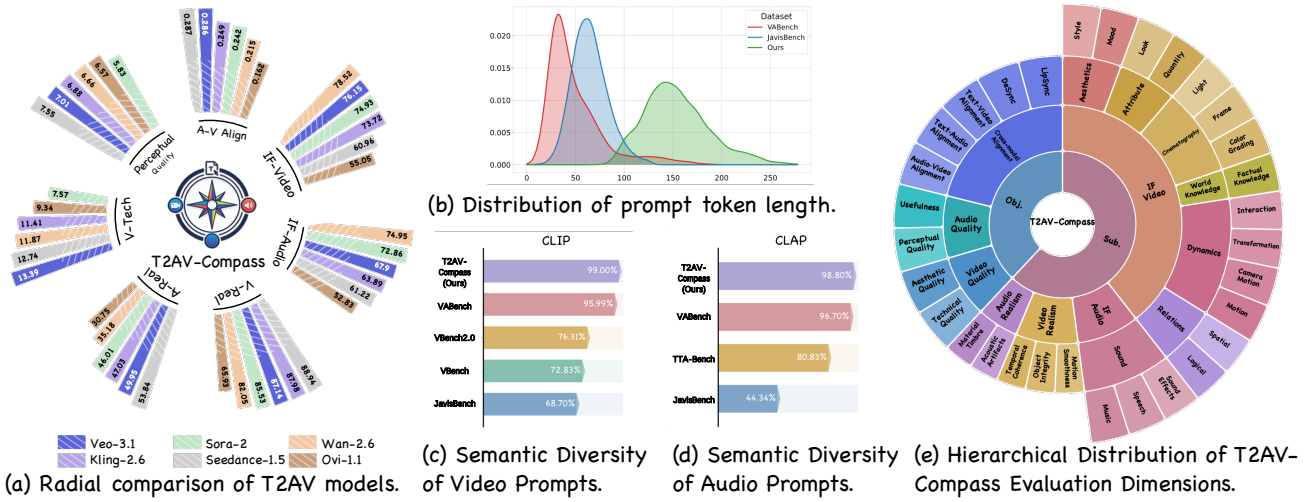

Figure 1. **Overview of T2AV-Compass analysis and evaluation taxonomy.** (a) Radial comparison of representative T2AV models under our evaluation suite. (b) Prompt token-length distribution. (c–d) Semantic diversity of video/audio prompts quantified via embedding similarity (higher indicates broader coverage). (e) Hierarchical distribution of evaluation dimensions, clearly organizing objective metrics and MLLM-based assessments across video, audio, and cross-modal alignment.

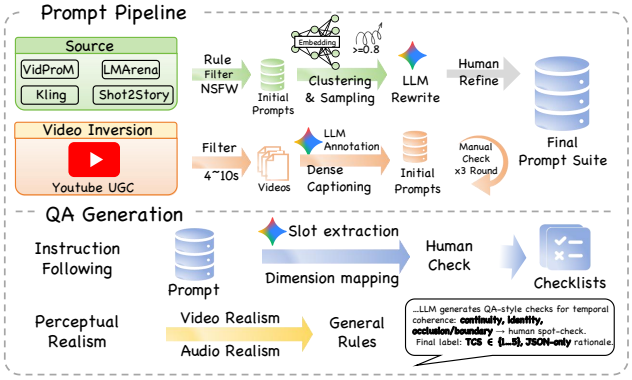

Figure 2. **Data construction and evaluation generation.** The prompt suite blends curated prompts (refined by LLMs) with a video-inversion stream. Finalized prompts are converted into dual-track checklists for instruction alignment and realism.

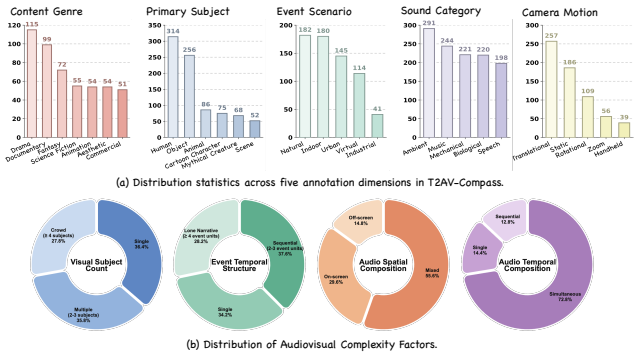

Figure 3. **Dataset statistics of T2AV-Compass.** (a) Category distributions across five key annotation dimensions. (b) Distributions of audiovisual complexity factors, including Visual Subject Count, Event Temporal Structure, Audio Spatial Composition, and Audio Temporal Composition.

ance, we encode prompts via Sentence-BERT [26], perform deduplication with a 0.8 cosine similarity threshold, and apply square-root sampling to preserve semantic distinctiveness while preventing the dominance of frequent topics.

Prompt Refinement. Since raw prompts often lack sufficient descriptive density for state-of-the-art models [6, 15, 24, 35], we employ Gemini-2.5-Pro to restructure and enrich them with explicit visual, dynamic, acoustic, and cinematographic constraints. A subsequent manual audit filters out static or illogical scenes, yielding a curated subset of **400 high-complexity prompts**.

Real-world Video Inversion. To counterbalance potential LLM hallucinations and ensure physical plausibility [3, 7], we introduce a Video-to-Text inversion stream. We select 100 diverse, high-quality YouTube clips (4–10s) and use Gemini-2.5-Pro to generate temporally aligned captions. Discrepancies between generated prompts and source content are resolved via human-in-the-loop verification, yielding 100 prompts anchored in real-world dynamics.

2.2. Dual-Level Evaluation Framework

To systematically evaluate T2AV models, we propose a dual-level framework, combining objective metrics with a reasoning-first MLLM-as-a-Judge protocol (Figure 4).

Objective Evaluation. We employ expert automated metrics across three pillars to establish a stable evaluation baseline. For **Video Quality**, we measure low-level technical fidelity via the **Video Technological Score (VT)** using DOVER++ [40] and high-level aesthetic appeal via the **Video Aesthetic Score (VA)** using Aesthetic Predictor V2.5 [2]. For **Audio Quality**, we employ reference-

Figure 4. **Illustration of the subjective evaluation framework.** Our protocol provides interpretable diagnosis through two tracks: *(Top)* Instruction following is evaluated via rigorous Q&A checklist pairs. *(Bottom)* Realism assesses perceptual quality, rewarding fine-grained details while penalizing visual hallucinations and audio dissonance.

free metrics based on Audiobox [32]: the **Audio Aesthetic Score (AA)** (averaging signal fidelity and semantic content usefulness) and the **Speech Quality Score (SQ)** via NISQA [23] to capture speech naturalness. For **Cross-modal Alignment**, we compute **Text–Audio (T–A)** and **Text–Video (T–V)** semantic consistency using CLAP [4] and VideoCLIP-XL-V2 [34], respectively. **Audio–Video (A–V)** alignment is evaluated via ImageBind [5]. Finally, **Temporal Synchronization** is quantified using **DeSync (DS)** (audio-visual onset offset via Synchformer [13]) and **LatentSync (LS)** [16] for talking-face lip-sync scenarios.

Subjective Instruction Following (IF). Traditional metrics often fail to capture complex cross-modal dynamics. We therefore employ Gemini-2.5-Pro as a judge to verify generated videos against QA checklists derived from each prompt. The IF taxonomy encompasses 7 primary dimensions: **Attribute** (look, quantity), **Dynamics** (motion, interaction, transformation, camera motion), **Cinematography** (lighting, framing, color grading), **Aesthetics** (style, mood), **Relations** (spatial, logical), **World Knowledge** (factual accuracy), and **Sound** (effects, speech, music). We enforce a *reasoning-first* protocol, requiring the MLLM to articulate a rationale before assigning a 1–5 score, which improves interpretability for error attribution.

Subjective Realism. While IF verifies prompt adherence, it may overlook internal inconsistencies or violations of physical plausibility. We therefore assess **Video Realism** through three complementary metrics: the **Motion Smoothness Score (MSS)** (penalizing jitter and temporal discontinuities), **Object Integrity Score (OIS)** (identifying anatomical/structural distortions), and **Temporal Coherence Score (TCS)** (assessing object permanence and occlusion logic). Additionally, **Audio Realism** is evaluated via the **Acoustic Artifacts Score (AAS)** (measuring background noise and mechanical distortion) and **Material–Timbre Consistency (MTC)** (verifying alignment between acoustic timbre and underlying physical properties).

3. Experiments

We evaluate 15 representative T2AV systems, comprising 7 closed-source end-to-end models (Veo-3.1 [6], Sora-2 [24], Kling-2.6 [15], Wan-2.6/2.5 [33], Seedance-1.5 [28], PixVerse-V5.5 [25]), 3 open-source end-to-end models (Ovi-1.1 [22], LTX-2 [9], JavisDiT [19]), and 5 composed pipelines (Wan-2.2/HunyuanVideo-1.5 [33, 39] combined with HunyuanVideo-Foley [29] or MMAudio [1], plus AudioLDM2+MTV [18, 38]). Table 1 reports the dual-level results and yields three core insights:

(1) Open vs. Closed-Source. Closed-source models dominate the leaderboard, with **Veo-3.1** ranking first in overall average (70.29), followed by Sora-2 (69.83), Kling-2.6 (68.16), and Wan-2.6 (67.68). Among open-source systems, **LTX-2** achieves the strongest overall performance

Table 1. Comprehensive evaluation of T2AV models. The table reports both **Objective metrics** (video quality, audio quality, and cross-modal alignment) and **Subjective metrics** (MLLM-based instruction following and realism evaluation). A dash (–) indicates that the model is unable to generate human speech.

Method	Open-Source	Objective Quality & Alignment									Subjective Evaluation (MLLM)				
		Video Quality		Audio Quality		Cross-modal Alignment					Instruction Following		Realism		Avg. ↑
		VT↑	VA↑	AA↑	SQ↑	A-V↑	T-A↑	T-V↑	DS↓	LS↑	IF Vid↑	IF Aud↑	Vid Real↑	Aud Real↑	
- T2AV (End-to-End)															
Veo-3.1	✗	13.39	5.425	6.818	1.597	0.2856	0.2335	0.2438	0.6776	1.509	76.15	67.90	87.14	49.95	70.29
Sora-2	✗	7.568	4.112	5.584	1.485	0.2419	0.2484	0.2432	0.8100	1.331	74.93	72.86	85.53	46.01	69.83
Kling-2.6	✗	11.41	5.417	6.666	1.783	0.2495	0.2495	0.2449	0.7852	1.502	73.72	63.89	87.98	47.03	68.16
Wan-2.6	✗	11.87	4.605	6.440	1.476	0.2149	0.2572	0.2451	0.8818	1.081	78.52	74.95	82.05	35.18	67.68
Seedance-1.5	✗	12.74	5.007	7.403	1.766	0.2875	0.2320	0.2370	0.8650	1.560	60.96	61.22	88.94	53.84	66.24
LTX-2	✓	7.160	4.661	6.742	1.597	0.1851	0.2365	0.2411	0.8756	1.339	63.97	64.74	89.95	36.23	63.72
Wan-2.5	✗	13.29	4.642	6.169	1.543	0.2026	0.2445	0.2470	0.8810	1.065	76.56	57.95	76.00	35.06	61.39
Pixverse-V5.5	✗	11.54	4.558	5.982	1.824	0.1816	0.2305	0.2431	0.6627	1.306	65.13	53.31	69.37	33.58	55.35
Ovi-1.1	✓	9.336	4.368	6.531	1.592	0.1620	0.1756	0.2391	0.9624	1.191	55.05	52.83	65.93	30.75	51.14
JavisDiT	✓	6.850	3.575	4.752	–	0.1284	0.1257	0.2320	1.322	–	32.56	15.26	34.97	14.85	24.41
- T2V + TV2A (Cascaded)															
HunyuanVideo1.5 + Hunyuan-Foley	✓	11.34	4.804	6.330	–	0.2598	0.2021	0.2436	0.8924	–	66.23	40.09	86.75	41.65	58.68
Wan-2.2 + Hunyuan-Foley	✓	13.43	5.605	6.353	–	0.2575	0.2076	0.2455	0.7935	–	64.54	37.10	89.63	41.25	58.13
Wan-2.2 + MMAudio	✓	13.43	5.605	6.076	–	0.2195	0.2448	0.2455	0.8890	–	64.79	38.19	89.63	36.05	57.17
HunyuanVideo1.5 + MMAudio	✓	11.34	4.804	6.101	–	0.2210	0.2466	0.2436	0.9427	–	66.10	35.94	85.38	35.15	55.64
- T2A + TA2V															
AudioLDM2 + MTV	✓	8.066	3.458	6.253	1.264	0.1639	0.2698	0.2394	1.1592	0.6835	47.13	54.39	56.73	31.90	47.54

(63.72) and the best **Video Realism** (89.95), while **Wan-2.6** leads in **Instruction Following** (IF-Vid 78.52, IF-Aud 74.95). Notably, the open/closed gap is most pronounced in high-level instruction adherence and audio realism, rather than in modality-specific quality alone in our benchmark.

(2) T2AV-Compass is Challenging. No single model demonstrates universal dominance across all dimensions. While Veo-3.1 attains the highest overall average, it still exhibits major deficiencies in **Audio Realism** (only 49.95), indicating a persistent "Audio Realism Bottleneck" across current systems, even among the strongest models..

(3) Cascaded Pipelines are Strong but Disjoint. Cascaded T2V → V2A systems match or exceed end-to-end models in modality-specific quality (e.g., **Wan-2.2 + Hunyuan-Foley** reaches 89.63 in Video Realism). However, they often lag in audio-visual alignment (e.g., DS and A-V metrics), suggesting limitations from fragmented optimization across stages.

Difficulty Trends. Figure 5 shows failure rates increase with prompt complexity, with the sharpest degradation on long-narrative prompts. This identifies long-horizon audio-visual generation as a key challenge for T2AV systems.

Human–MLLM Judge Agreement Analysis. On a 50-prompt subset, Gemini 2.5 Pro shows the closest agreement with human ratings overall, while **Audio Realism** remains harder and benefits from additional human verification.

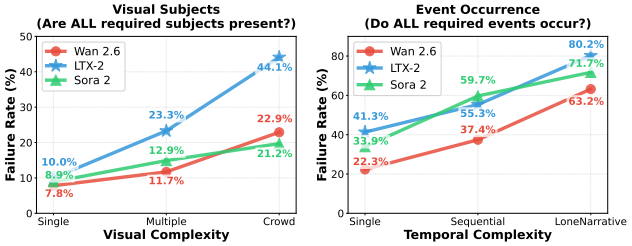

Figure 5. **Difficulty trends.** Failure rates are reported by prompt complexity, grouped by *Visual Subject Count* and *Event Temporal Structure* (lower is better).

Table 2. **L1 distance** between evaluators across four subjective dimensions (lower is better).

Evaluator	IF Video↑	IF Audio↑	Video Real.↑	Audio Real.↑	Overall↑
Inter-Human	0.917	0.911	0.926	1.042	0.949
Gemini-2.5-Pro	1.012	**0.980**	**0.937**	1.420	1.087
Gemini-2.5-Flash	1.212	1.027	1.397	**1.193**	1.207
Qwen3-Omni-Flash	1.026	1.887	1.297	1.680	1.473

4. Conclusion

We present **T2AV-Compass**, a unified benchmark for T2AV evaluation, with a taxonomy-driven prompt pipeline and a dual-level framework combining objective metrics and MLLM-based judging. Across 15 systems, experiments reveal a consistent gap: despite strong visual fidelity, current models remain limited in audio realism and long-horizon prompt following.

References

[1] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28901–28911, 2025. 3

[2] discus0434. Aesthetic Predictor V2.5: High-resolution image aesthetics prediction, 2023. 2

[3] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025. 2

[4] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[5] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, 2023. 3

[6] Google DeepMind. Veo: Generative video models. https://deepmind.google/technologies/veo, 2024. 1, 2, 3

[7] Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2VPhysBench: A first-principles benchmark for physical consistency in text-to-video generation. *arXiv preprint arXiv:2505.00337*, 2025. 2

[8] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1

[9] Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026. 3

[10] Mingfei Han, Linjie Yang, Xiaojun Chang, Lina Yao, and Heng Wang. Shot2Story: A new benchmark for comprehensive understanding of multi-shot videos. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[11] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen Video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[12] 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 (CVPR)*, pages 21807–21818, 2024. 1

[13] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10270–10281, 2023. 3

[14] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 119–132. Association for Computational Linguistics, 2019. 1

[15] Kuaishou Technology. Kling AI: Text-to-video generation service. Online service, 2024. 1, 2, 3

[16] Chunyu Li, Chao Zhang, Weikai Xu, Jingyu Lin, Jinghui Xie, Weiguo Feng, Bingyue Peng, Cunjian Chen, and Weiwei Xing. LatentSync: Taming audio-conditioned latent diffusion models for lip sync with syncnet supervision. *arXiv preprint arXiv:2412.09262*, 2024. 3

[17] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. VideoDirectorGPT: Consistent multi-scene video generation via llm-guided planning, 2023. 1

[18] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024. 1, 3

[19] Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Rongxin Jiang, Jiebo Luo, Hao Fei, et al. JavisDiT: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv preprint arXiv:2503.23377*, 2025. 3

[20] 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 (CVPR)*, pages 22139–22149, 2024. 1

[21] LMArena Community. LMArena: Open arena for evaluating large multimodal models. https://lmarena.ai/, 2024. Online benchmarking and prompt collection platform. 1

[22] Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025. 3

[23] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. NISQA: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Interspeech 2021*, pages 2127–2131, 2021. 3

[24] OpenAI. Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators, 2024. 1, 2, 3

[25] PixVerse Team. PixVerse V5.5: Ai video generation with built-in sound and multi-shot scenes. https://app.pixverse.ai/, 2025. 3

[26] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics, 2019. 2

[27] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. MM-Diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10219–10228, 2023. 1

[28] Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025. 3

[29] Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. HunyuanVideo-Foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. *arXiv preprint arXiv:2508.16930*, 2025. 3

[30] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-video generation without text-video data. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 1

[31] Xiaoxuan Tang, Xinping Lei, Chaoran Zhu, Shiyun Chen, Ruibin Yuan, Yizhi Li, Changjae Oh, Ge Zhang, Wenhao Huang, Emmanouil Benetos, et al. AutoMV: An automatic multi-agent system for music video generation. *arXiv preprint arXiv:2512.12196*, 2025. 1

[32] Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023. 3

[33] 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. 3

[34] Jiapeng Wang, Chengyu Wang, Kunzhe Huang, Jun Huang, and Lianwen Jin. VideoCLIP-XL: Advancing long description understanding for video clip models. *arXiv preprint arXiv:2410.00741*, 2024. 3

[35] Jun Wang, Xijuan Zeng, Chunyu Qiang, Ruilong Chen, Shiyao Wang, Le Wang, Wangjing Zhou, Pengfei Cai, Jiahui Zhao, Nan Li, et al. Kling-Foley: Multimodal diffusion transformer for high-quality video-to-audio generation. *arXiv preprint arXiv:2506.19774*, 2025. 2

[36] Qunzhong Wang, Jie Liu, Jiajun Liang, Yilei Jiang, Yuanxing Zhang, Jinyuan Chen, Yaozhi Zheng, Xintao Wang, Pengfei Wan, Xiangyu Yue, et al. VR-Thinker: Boosting video reward models through thinking-with-image reasoning, 2025. 1

[37] Wenhao Wang and Yi Yang. VidProM: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *Advances in Neural Information Processing Systems*, 37: 65618–65642, 2024. 1

[38] Shuchen Weng, Haojie Zheng, Zheng Chang, Si Li, Boxin Shi, and Xinlong Wang. Audio-sync video generation with multi-stream temporal control. *arXiv preprint arXiv:2506.08003*, 2025. 3

[39] Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. HunyuanVideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025. 3

[40] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang, Wenxiu Sun Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20144–20154, 2023. 2

[41] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. 1