JWB-DH-V1: Benchmark for Joint Whole-Body Talking Avatar and Speech Generation Version I

Xinhan Di[1*] ✉ Kristin Qi[2*] Pengqian Yu[3]
[1]Independent Researcher, China
[2]Computer Science, University of Massachusetts Boston, USA
[3]National University of Singapore, Singapore
deepearthgo@gmail.com, yanankristin.qi001@umb.edu, yupengqian1989@gmail.com

Abstract

Recent advances in diffusion-based video generation have enabled photo-realistic short clips, but current methods still struggle to achieve multi-modal consistency when jointly generating whole-body motion and natural speech. Current approaches lack comprehensive evaluation frameworks that assess both visual and audio quality, and there are insufficient benchmarks for region-specific performance analysis. To address these gaps, we introduce the Joint Whole-Body Talking Avatar and Speech Generation Version I(JWB-DH-V1), comprising a large-scale multi-modal dataset with 10,000 unique identities across 2 million video samples, and an evaluation protocol for assessing joint audio-video generation of whole-body animatable avatars. Our evaluation of SOTA models reveals consistent performance disparities between face/hand-centric and whole-body performance, which incidates essential areas for future research. The dataset and evaluation tools are publicly available at https://github.com/deepreasonings/WholeBodyBenchmark.

1. Introduction

Diffusion-based video generation has emerged as a powerful approach for synthesizing realistic videos [31]. Researchers are actively exploring cascaded and latent-space diffusion pipelines to address existing limitations [19, 38, 42]. However, in the context of digital humans [24, 27], current diffusion models still struggle to synthesize realistic full-body motion and appearance effectively [4, 5, 8, 13, 23, 35]. These models rarely achieve synchronized audio and video generation for full-body speakers. This gap highlights the need for integrated systems capable of producing high-quality full-body motion synchronized naturally with speech, which is the focus of the Joint Whole-Body Talking Avatar and Speech Generation Dataset.

Moreover, creating fully natural whole-body talking avatars remains a substantial challenge. Advanced talking head models generally concentrate on facial and upper-body regions [4, 5, 8, 13] and fail to synthesize full-body motions or gestures aligned with speech [23, 35]. Furthermore, existing methods typically do not achieve stable and joint audio-video generation from a single frame [6]. These limitations reveal the importance of establishing a benchmark specifically designed to evaluate and advance methods that integrate facial animation with whole-body motion, ensuring coherence with the generated speech audio.

2. Related Work

2.1. Diffusion-Based Video Generation and Talking Head Generation

Diffusion-based models have been extended from images to video [9], with Stable Video Diffusion as a foundation [2]. Later works incorporate conditioning and multi-scale processing [11, 16, 29, 41], but lack audio-visual alignment for human-centric generation. Lip-sync discriminators guide GANs or diffusion models for facial animation [24, 27, 44], yet these approaches are spatially constrained to head or upper body [5, 8, 13, 14, 18], rather than enabling full-body audio-driven generation.

2.2. Joint Video-Audio Generation

Recent advances video and audio synthesis include DeepMind's Veo-3 [6] and unified Diffusion Transformers [21, 43]. These systems enable synchronized multi-modal content creation. However, current joint audio-video generation approaches primarily emphasize general alignment, rather than fine-grained synchronization between speech and body movements, which is essential for digital human applications. Moreover, achieving the stable joint audio-video generation from a single input frame remains experimental.

* These authors contributed equally to this work.

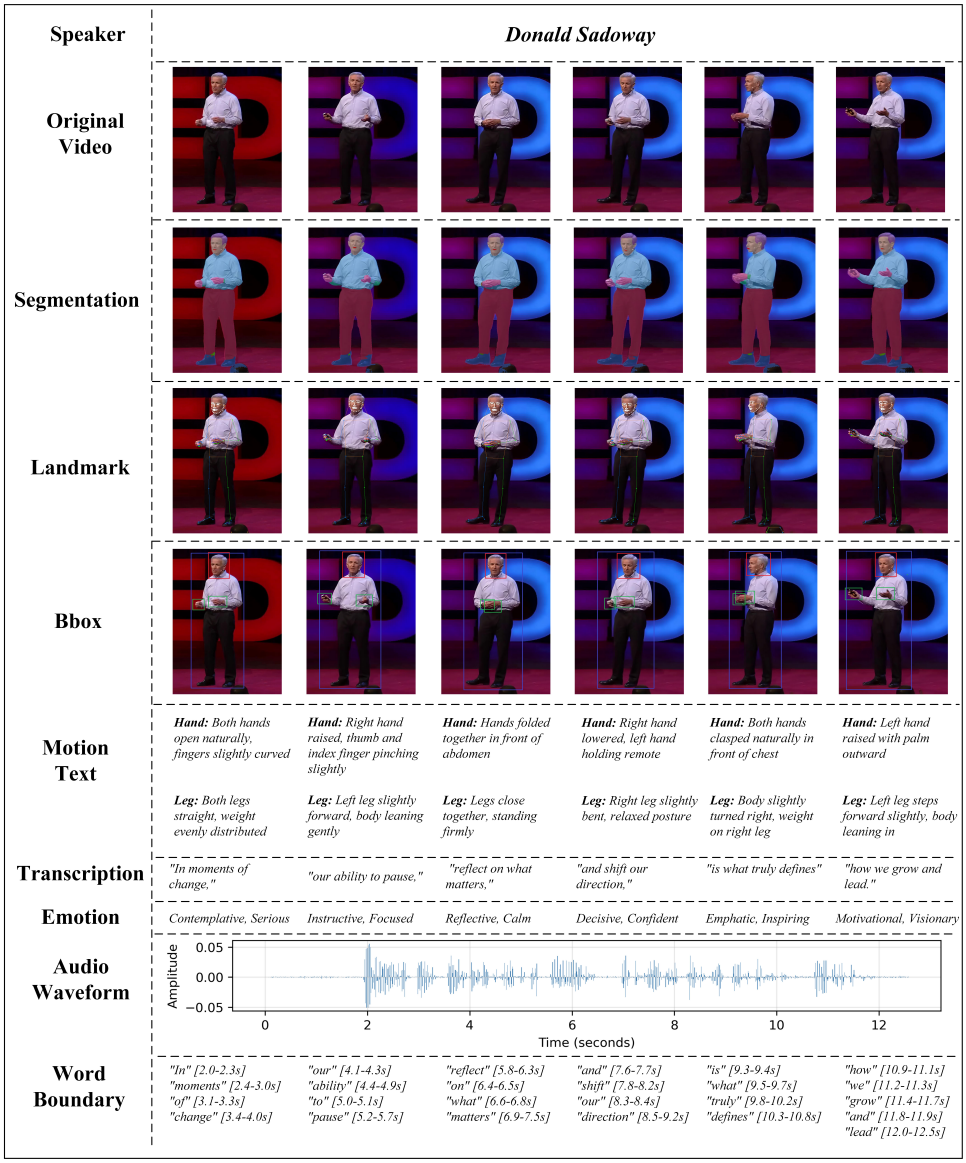

Figure 1. Illustration of our dataset annotations. Each column represents a key frame. From top to bottom: the original frame, body segmentation, landmark, bounding box annotations for key regions (hands, legs, whole body), motion text describing pose semantics, corresponding speech transcription, the motion flag of each transcription, ground truth audio and word boundary. The dataset captures fine-grained multi-modal alignment between body postures, hand gestures, leg stance, and spoken language with its audio over time.

To address these challenges, we introduce the Whole-Body Joint Talking Avatar and Speech Generation benchmark that is designed to advance research in synchronized and high-quality audio-video generation.

3. Method

3.1. Dataset Curation

We curate the dataset with **10,000 unique identities**, each appearing in approximately **200 different scene configurations**, totaling **2 million samples of video clips**. The annotation for each sample is demonstrated in Figure 1. **20,000 samples** are used for evaluation.

3.2. Video Generation Sub Evaluation Protocol

To evaluate perceptual quality and temporal stability of generated whole-body avatars, we adopt six reference-free video metrics: (1) Subject Consistency (SC): temporal appearance consistency via DINO features [3], (2) Background Consistency (BC): background stability via CLIP features [30], (3) Motion Smoothness (MS): optical flow continuity [20], (4) Dynamic Degree (DD): motion rich-

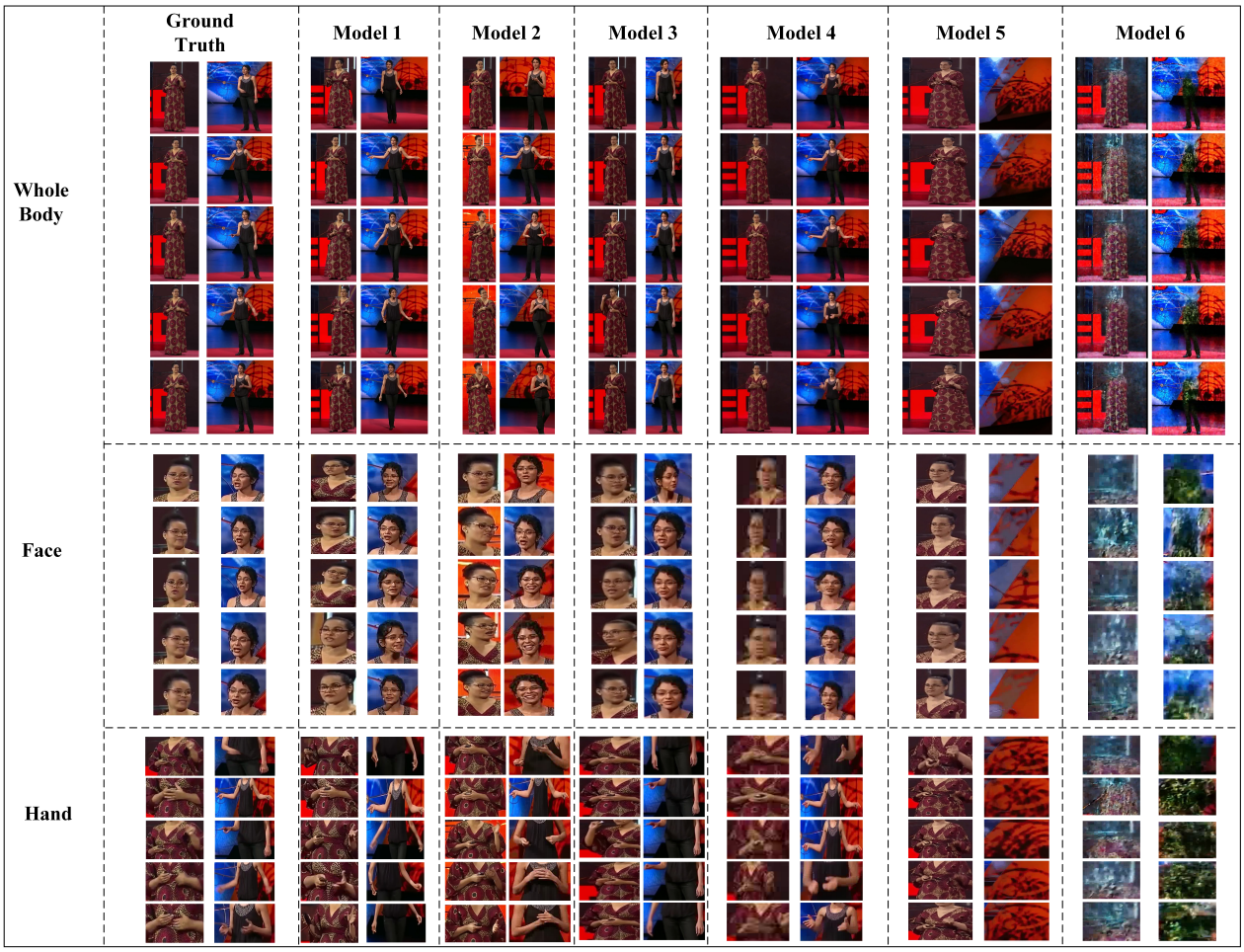

Figure 2. Visual comparison across full body, face, and hand regions. Model 1(Wan [37]),2(OpenS [26]),3(Hun [16]), 4(ecv2/w [23]) show good performance in structure, motion, and identity preservation. Model 5(ecv2/wo [23]),6(Ha3/wo [5]) show noticeable artifacts and degradation across all regions.

Table 1. Comparison of different text-to-speech (TTS) models on a range of speech synthesis tasks.

Model	Emotions		Paralinguistics		Complex Pronunciation		Syntactic Complexity	
	WER	Win-Rate	WER	Win-Rate	WER	Win-Rate	WER	Win-Rate
gpt-4o-mini-tts (baseline)	5.76	-	24.59	-	**28.92**	-	6.09	-
Suno Bark [32]	6.34	0.00%	29.83	6.67%	55.81	8.42%	6.02	15.29%
Tortoise-TTS [1]	15.89	3.71%	59.72	5.08%	59.84	5.47%	5.13	17.42%
MiniCPM [10]	17.62	8.49%	26.53	6.45%	61.12	8.11%	8.83	16.28%
Qwen 2.5 Omni [40]	8.47	31.13%	23.06	24.02%	47.69	13.37%	9.01	28.63%
11Labs eleven multilingual v2 [7]	**4.68**	20.38%	22.56	35.41%	33.49	14.83%	9.12	33.92%
gpt-4o-mini-audio-preview [25]	5.79	58.31%	24.97	60.68%	35.58	37.63%	10.81	53.72%

ness from flow magnitude [33], (5) Aesthetic Quality (AQ): LAION aesthetic score [17], and (6) Imaging Quality (IQ): visual clarity via MUSIQ [15]. These metrics provide a comprehensive evaluation of realism and coherence.

3.3. Co-Speech Sub Evaluation Protocol

To complement the region-level evaluation of the whole-body generated avatars, we adopt six standard metrics to assess frame quality, temporal coherence, and identity preservation [4, 5, 8, 13]. These include FID [28], FVD [36], SSIM [39], PSNR [12], E-FID [28], and CSIM [28]. These metrics capture fidelity, temporal stability, and identity consistency for comprehensive evaluation.

Table 2. Evaluation of eight models across three regions and twelve metrics. SC: Subject Consistency, BC: Background Consistency, MS: Motion Smoothness, DD: Dynamic Degree, AQ: Aesthetic Quality, IQ: Imaging Quality, FID/FVD: Fréchet (Video) Distance, SSIM: Structural Similarity, PSNR: Peak Signal-to-Noise Ratio, E-FID: Enhanced FID, CSIM: Cosine Similarity. **Bold** = best among generative models (ground-truth excluded).

Region	Metric	GT	Step[11]	Hun[16]	Wan[37]	OpenS[26]	Ha3/w[5]	Ha3/wo[5]	ECV2/w[23]	ECV2/wo[23]
Whole Body	SC	100	93.45	92.54	**96.83**	97.12	20.20	10.74	29.16	7.08
	BC	100	94.50	94.30	**96.31**	98.00	19.91	10.93	31.25	6.49
	MS	100	98.71	98.50	**99.69**	90.69	20.60	11.37	29.75	7.14
	DD	100	67.39	**69.34**	59.89	60.43	4.65	4.00	9.81	0.09
	AQ	100	**48.32**	43.17	46.73	32.88	10.01	3.84	14.27	2.52
	IQ	100	61.99	**63.54**	54.67	42.57	9.88	5.09	16.53	3.29
	FID	0.0	172.47	208.75	**92.76**	121.53	491.43	568.68	508.19	572.14
	FVD	0.0	1356.49	1567.94	**750.51**	896.94	2470.14	4366.39	2519.77	3791.25
	SSIM	1.0	0.573	0.543	**0.660**	0.623	0.209	0.016	0.216	0.028
	PSNR	∞	17.74	16.06	**20.16**	19.48	6.24	0.83	6.17	1.29
	E-FID	0.0	**187.58**	216.91	196.39	218.78	484.08	634.60	492.17	712.91
	CSIM	1.0	0.760	0.811	**0.896**	0.876	0.303	0.053	0.299	0.092
Face	SC	100	65.06	55.07	**66.51**	52.34	14.92	12.23	14.79	8.29
	BC	100	**74.04**	68.05	72.26	63.55	15.73	15.15	18.16	9.27
	MS	100	63.12	56.78	**63.40**	49.74	14.66	12.15	12.73	8.16
	DD	100	14.91	4.15	**25.12**	7.69	7.21	4.55	2.42	0.21
	AQ	100	**28.18**	21.26	25.09	19.76	5.79	2.99	5.71	5.69
	IQ	100	**39.40**	34.19	33.68	30.90	6.98	4.60	8.49	6.29
	FID	0.0	300.98	291.21	**254.03**	275.50	425.82	720.22	437.27	801.27
	FVD	0.0	1811.92	1351.71	1591.22	**1221.85**	2794.85	3255.19	2651.92	3047.27
	SSIM	1.0	0.372	0.322	**0.417**	0.388	0.127	0.010	0.139	0.027
	PSNR	∞	13.54	14.47	**16.99**	16.29	6.75	0.89	6.88	1.25
	E-FID	0.0	339.18	347.24	**273.97**	359.85	451.76	492.67	439.92	513.58
	CSIM	1.0	0.736	0.787	**0.866**	0.863	0.385	0.065	0.401	0.062
Hand	SC	100	**66.83**	48.15	51.27	49.16	13.65	2.13	14.19	3.58
	BC	100	**77.98**	64.98	57.45	32.45	14.76	2.56	18.21	4.27
	MS	100	46.36	**49.65**	42.49	43.30	13.25	0.49	9.27	2.79
	DD	100	48.14	**52.85**	41.71	43.89	13.05	0.91	9.52	1.84
	AQ	100	**24.22**	19.02	18.38	16.40	5.08	0.39	5.08	2.74
	IQ	100	**38.40**	27.69	23.84	21.50	6.51	0.89	10.24	1.94
	FID	0.0	327.10	301.94	331.51	**278.95**	574.40	921.16	588.72	975.92
	FVD	0.0	1597.65	1734.98	1561.58	**1534.50**	7750.48	10671.79	8958.27	9472.28
	SSIM	1.0	0.316	0.262	**0.367**	0.342	0.098	0.008	0.106	0.019
	PSNR	∞	12.98	11.45	13.09	**15.50**	6.13	0.55	7.27	0.49
	E-FID	0.0	277.47	255.53	**240.48**	264.98	445.59	713.46	525.18	741.03
	CSIM	1.0	0.681	0.737	**0.796**	0.794	0.366	0.041	0.427	0.152

3.4. Speech Audio Sub Evaluation Protocol

To evaluate the quality of generated audio, we use Large-Audio-Language Models (LALMs) as outlined in recent evaluation methods [22]. It targets aspects of speech synthesis such as prosody, pausing, expressiveness, and pronunciation that are not adequately captured by traditional metrics like Word Error Rate (WER). Therefore, alongside WER, we select Gemini 2.5 Pro [34] as our primary LALM-based evaluator [22]. To evaluate a candidate text-to-speech (TTS) system T_i, we compare it against a strong reference system T_j that has low WER to ensure high-fidelity synthesis of the benchmark utterances. We adopt a win-rate-based metric to summarize performance. Let $W(T_i)$ denote the win-rate of system T_i relative to the baseline T_j. $W(T_i)$ is computed as: $W(T_i) = \frac{\sum(\text{winner}=\text{index}_i)+0.5\cdot\sum(\text{winner}=0)}{n}$, where index $x_i \in \{1, 2\}$ corresponds to the randomized label assigned to T_i, and n denotes the total number of comparisons.

4. Initial Evaluation

We evaluate four groups of state-of-the-art (SOTA) models: (1) open-source video generation [11, 16, 26, 37], (2) open-source talking avatars [5, 23], (3) open-source TTS systems [1, 10, 32, 40], and (4) closed-source TTS systems [7, 25]. Table 1 and 2 report audio and video metrics performance, respectively. Figure 2 shows region-specific results for the face, hands, and full body. The joint audio-video model [6] is excluded due to instability in generating synchronized outputs from a single frame.

Video generation models take an initial video frame and a text prompt to guide motions (Figure 1). Talking avatar models are driven by audio with or without pose guidance, which are referred to as Ha3/w and Ha3/wo [5]. To evaluate regional generative quality, we evaluate face, hands, and whole body independently. Additionally, utterance texts with prompting strategies (e.g., emotion prompts in Figure 1) are used to evaluate each TTS system.

5. Conclusion

We present JWB-DH-V1 (Joint Whole-Body Talking Avatar and Speech Generation Benchmark Version 1), a large-scale public dataset with evaluation protocols for evaluating joint whole-body avatar generation. Future releases will support 60-second clips with refined audio-video annotations and will further evaluate SOTA stable joint generation models from a single initial frame.

References

[1] J. Betker. Better Speech Synthesis Through Scaling: Tortoise, an expressive multi-voice tts system. *arXiv preprint arXiv:2305.07243*, May 2023. URL https://arxiv.org/abs/2305.07243. Introduces TorToiSe, combining autoregressive and diffusion models for high-quality multi-voice TTS; code released under Apache-2.0 :contentReference[oaicite:1]index=1. 3, 4

[2] 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. 1

[3] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[4] J. Cui, H. Li, Y. Yao, H. Zhu, H. Shang, K. Cheng, H. Zhou, S. Zhu, and J. Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*, 2024. 1, 3

[5] J. Cui, H. Li, Y. Zhan, H. Shang, K. Cheng, Y. Ma, S. Mu, H. Zhou, J. Wang, and S. Zhu. Hallo3: Highly dynamic and realistic portrait image animation with diffusion transformer networks. *arXiv preprint arXiv:2412.00733*, 2024. 1, 3, 4

[6] G. DeepMind. Veo 3: Video, Meet Audio. Tech blog, 2025. First video generator with native audio generation. 1, 4

[7] ElevenLabs. Eleven Multilingual v2: A foundational multilingual text-to-speech model supporting 29 languages. Official ElevenLabs blog, Aug. 2023. Released out of beta August22,2023; supports emotionally rich AI voice in 29 languages while preserving speaker identity :contentReference[oaicite:1]index=1. 3, 4

[8] J. Guan, Z. Zhang, H. Zhou, T. Hu, K. Wang, D. He, H. Feng, J. Liu, E. Ding, Z. Liu, et al. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *Proceedings of the IEEE/CVF CVPR*, pages 1505–1515, 2023. 1, 3

[9] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 1

[10] S. Hu, Y. Tu, X. Han, C. He, G. Cui, X. Long, Z. Zheng, Y. Fang, Y. Huang, W. Zhao, X. Zhang, Z. Zhai, N. Ding, C. Jia, G. Zeng, D. Li, Z. Liu, and M. Sun. MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies. *arXiv preprint arXiv:2404.06395*, Apr. 2024. URL https://arxiv.org/abs/2404.06395. Introduces MiniCPM, including 1.2B and 2.4B-parameter SLMs with scalable training recipes and performance competitive with 7B–13B LLMs. 3, 4

[11] H. Huang, G. Ma, N. Duan, X. Chen, C. Wan, R. Ming, T. Wang, B. Wang, Z. Lu, A. 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, 4

[12] Q. Huynh-Thu and M. Ghanbari. The accuracy of psnr in predicting video quality for different video scenes and frame rates. *Telecommunication systems*, 49:35–48, 2012. 3

[13] X. Ji, X. Hu, Z. Xu, J. Zhu, C. Lin, Q. He, J. Zhang, D. Luo, Y. Chen, Q. Lin, et al. Sonic: Shifting focus to global audio perception in portrait animation. *arXiv preprint arXiv:2411.16331*, 2024. 1, 3

[14] J. Jiang, C. Liang, J. Yang, G. Lin, T. Zhong, and Y. Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. *arXiv preprint arXiv:2409.02634*, 2024. 1

[15] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 3

[16] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 3, 4

[17] LAION-AI. Laion aesthetic predictor. https://github.com/LAION-AI/aesthetic-predictor, 2022. Accessed: 2025-04-21. 3

[18] C. Li, C. Zhang, W. Xu, J. Xie, W. Feng, B. Peng, and W. Xing. Latentsync: Audio conditioned latent diffusion models for lip sync. *arXiv preprint arXiv:2412.09262*, 2024. 1

[19] X. Li, W. Chu, Y. Wu, W. Yuan, F. Liu, Q. Zhang, F. Li, H. Feng, E. Ding, and J. Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023. 1

[20] Z. Li, Z.-L. Zhu, L.-H. Han, Q. Hou, C.-L. Guo, and M.-M. Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023. 2

[21] K. Liu, W. Li, L. Chen, S. Wu, Y. Zheng, J. Ji, F. Zhou, R. Jiang, J. Luo, H. Fei, and T. Chua. Javis-DiT: Joint audio–video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv preprint arXiv:2503.23377*, Mar. 2025. 1

[22] R. R. Manku, Y. Tang, X. Shi, M. Li, and A. Smola. Emergenttts-eval: Evaluating tts models on complex prosodic, expressiveness, and linguistic challenges using model-as-a-judge. *arXiv preprint arXiv:2505.23009*, May 2025. URL https://arxiv.org/abs/2505.23009. Introduces the EmergentTTS-Eval benchmark and a model-as-a-judge evaluation framework using a Large Audio Language Model (LALM). 4

[23] R. Meng, X. Zhang, Y. Li, and C. Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. *arXiv preprint arXiv:2411.10061*, 2024. 1, 3, 4

[24] D. Min, M. Song, E. Ko, and S. J. Hwang. Styletalker: One-shot style-based audio-driven talking head video generation. *arXiv preprint arXiv:2208.10922*, 2022. 1

[25] OpenAI & Microsoft Azure AI Foundry. GPT-4o-Mini-Audio-Preview: A compact, cost-efficient audio-capable

multimodal model. Azure AI Services / OpenAI Platform documentation, Dec. 2024. Public preview of audio-in/audio-out capabilities via Chat Completions API (model version 2024-12-17); 25 3, 4

[26] X. Peng, Z. Zheng, and C. S. *et al.* Open-sora 2.0: Training a commercial-level video generation model in 200k. *arXiv preprint arXiv:2503.09642*, 2025. 3, 4

[27] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *the 28th ACM international conference on multimedia*, pages 484–492, 2020. 1

[28] Z. Qin, R. Zheng, Y. Wang, T. Li, Z. Zhu, M. Yang, M. Yang, and L. Wang. Versatile multimodal controls for whole-body talking human animation. *arXiv preprint arXiv:2503.08714*, 2025. 3

[29] D. Qiu, Z. Fei, R. Wang, J. Bai, C. Yu, M. Fan, G. Chen, and X. Wen. Skyreels-a1: Expressive portrait animation in video diffusion transformers. *arXiv preprint arXiv:2502.10841*, 2025. 1

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[31] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1

[32] Suno AI. Bark: Text-prompted generative audio model. GitHub and Hugging Face repositories, Apr. 2023. Released under MIT license in April/2023; transformer-based model capable of generating realistic multilingual speech, music, sound effects, and nonverbal audio. 3, 4

[33] Z. Teed and J. Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer. 3

[34] The Economic Times. Google introduces stable gemini 2.5 flash and pro, previews gemini 2.5 flash-lite. *The Economic Times (India)*. Announces general availability of Gemini2.5Pro alongside Gemini2.5Flash and preview of Flash-Lite. 4

[35] S. Tu, Z. Xing, X. Han, Z.-Q. Cheng, Q. Dai, C. Luo, and Z. Wu. Stableanimator: High-quality identity-preserving human image animation. *arXiv preprint arXiv:2411.17697*. 1

[36] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 3

[37] A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, J. Zeng, J. Wang, J. Zhang, J. Zhou, J. Wang, J. Chen, K. Zhu, K. Zhao, K. Yan, L. Huang, M. Feng, N. Zhang, P. Li, P. Wu, R. Chu, R. Feng, S. Zhang, S. Sun, T. Fang, T. Wang, T. Gui, T. Weng, T. Shen, W. Lin, W. Wang, W. Wang, W. Zhou, W. Wang, W. Shen, W. Yu, X. Shi, X. Huang, X. Xu, Y. Kou, Y. Lv, Y. Li, Y. Liu, Y. Wang, Y. Zhang, Y. Huang, Y. Li, Y. Wu, Y. Liu, Y. Pan, Y. Zheng, Y. Hong, Y. Shi, Y. Feng, Z. Jiang, Z. Han, Z.-F. Wu, and Z. Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 4

[38] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *IJCV*, pages 1–20, 2024. 1

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612, 2004. 3

[40] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, B. Zhang, X. Wang, Y. Chu, and J. Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, Mar. 2025. URL https://arxiv.org/abs/2503.20215. Describes Qwen2.5-Omni, an end-to-end streaming multimodal model (text, image, audio, video in/out) with Thinker-Talker architecture and TMRoPE, achieving state-of-the-art performance on OmniBench and speech generation benchmarks. 3, 4

[41] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1

[42] S. Zhang, J. Wang, Y. Zhang, K. Zhao, H. Yuan, Z. Qin, X. Wang, D. Zhao, and J. Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 1

[43] L. Zhao, L. Feng, D. Ge, F. Yi, C. Zhang, X. Zhang, and X. Li. UniForm: A unified diffusion transformer for audio–video generation. *arXiv preprint arXiv:2502.03897*, Feb. 2025. 1

[44] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021. 1