WAN-S2V: AUDIO-DRIVEN CINEMATIC VIDEO GENERATION

HumanAIGC Team
Tongyi Lab, Alibaba

ABSTRACT

Current state-of-the-art (SOTA) methods for audio-driven character animation demonstrate promising performance for scenarios primarily involving speech and singing. However, they often fall short in more complex film and television productions, which demand sophisticated elements such as nuanced character interactions, realistic body movements, and dynamic camera work. To address this long-standing challenge of achieving film-level character animation, we propose an audio-driven model, which we refere to as Wan-S2V, built upon Wan. Our model achieves significantly enhanced expressiveness and fidelity in cinematic contexts compared to existing approaches. We conducted extensive experiments, benchmarking our method against cutting-edge models such as Hunyuan-Avatar and Omnihuman. The experimental results consistently demonstrate that our approach significantly outperforms these existing solutions. Additionally, we explore the versatility of our method through its applications in long-form video generation and precise video lip-sync editing.

1 INTRODUCTION

Audio-driven human video generation has made significant progress recently, largely thanks to the development and application of Diffusion models Ho et al. (2020). Beginning with UNet-based text-to-image models and progressing to the latest DiT-based text-to-video models Wan et al. (2025); Kong et al. (2025), the quality of visual generation has also dramatically improved. Consequently, audio-driven models leveraging these latest DiT-based video foundation models are garnering increasing research attention Lin et al. (2025); Chen et al. (2025); Wang et al. (2025). However, influenced by prior work, current research primarily confines audio-driven models to single-scene human video generation, or even solely to single-character video driving. Nevertheless, in more complex scenarios such as film and television productions or multi-person scenes, audio-driven models still face numerous challenges. For instance, key questions arise: How can audio control a character while ensuring their movements are consistent and coordinated with the overall scene? How can person interactions be managed by audio and prompt jointly? This paper primarily focuses on audio-driven human generation in such complex scenarios as film and television, aiming to enhance the efficacy of audio-driven generation through comprehensive data acquisition, robust model training, and clever yet effective inference strategies.

Achieving film-quality audio-driven video, we contend, requires simultaneously leveraging the distinct yet complementary capabilities of text and audio. From a practical user perspective, text is optimally utilized for delineating the overarching dynamics of the video, including cinematic camera movements, comprehensive character trajectories, and interactions between entities. Audio, conversely, excels at dictating minute details such as character expressions and localized actions, including precise hand gestures and head orientation. Firstly, we construct our audio-driven model by leveraging the latest Wan text-to-video foundation modelWan et al. (2025). Our aim is to integrate audio-driven capabilities while preserving its inherent text control. Crucially, to ensure our model maintains text-control fidelity during training, we utilized Qwen-VL's Bai et al. (2025) video understanding capabilities for detailed textual captioning of videos, with a particular emphasis on descriptions pertinent to character motion. To effectively support generation in complex scenarios, such as film and television productions, we curated film and television-related audio-visual data from existing open-source datasets and augmented it with our own internally collected dataset of talking and singing character videos to form our comprehensive training dataset. While some exist-

ing methods attempt to reduce training complexity by training only partial network parameters, this often leads to conflicts between text and audio control. We hypothesize that a larger model capacity is more conducive to learning superior and harmonious text and audio control, thereby mitigating such conflicts. To facilitate large-scale, full-parameter training, and drawing inspiration from established parallel training paradigms for video foundation models, we implemented a hybrid training strategy combining FSDP Zhao et al. (2023) with Context Parallel, significantly accelerating the training process. Furthermore, to ensure enhanced stability and performance, we employed a multi-stage training regimen. This includes pre-training of the audio processing modules, followed by a comprehensive pre-training phase on the entire dataset, and subsequent fine-tuning on high-quality data. Collectively, these systematic strategies enable us to develop a robust and efficient audio-driven human video generation model.

Long video generation is crucial for generating videos in film and television scenarios. However, it faces challenges in maintaining stable details and consistency in scenes and even motion. Audio-driven methods like Tian et al. (2024) have attempted to use Motion Frames to maintain consistency between multiple clips, but an excessive number of motion frames can drastically increase computational complexity. This leads to a relatively limited number of motion frames, making it difficult to maintain long-term video stability in film and television scenarios. To address this, we introduce a approach similar to Zhang & Agrawala (2025), which effectively reduces the token count of Motion Frames by employing different token compression ratios at different times. This ensures the incorporation of more Motion Frames, thereby enabling the generation of more stable long videos.

To train our model, we constructed a dataset containing over clips, based on both publicly available video datasets and our own collected video data. This comprehensive dataset includes videos from solo scenarios focusing on human speech and singing, as well as complex character videos from film and television dramas.

Our main contributions are as follows:

- **Extending Audio-Driven Generation to Complex Scenarios:** We go beyond talking heads by enabling the creation of natural and expressive character movements in diverse and challenging scenes, incorporating both text-guided global motion control and audio-driven fine-grained local motion.
- **Long Video Stabilization and Efficient Model Variants:** We tackle the challenges of long video generation through optimized motion frame token reduction.
- **Comprehensive Training Data**We leverage a large-scale, diverse dataset to train our model and validate the effectiveness of our model through extensive experiments.

2 Data Processing Pipeline

Data Collection. Human-driven narratives constitute the core element of video content. Our objective is to identify videos featuring one or more human characters engaged in specific activities. Specifically, we adopted a two-pronged strategy:

Automated screening of large-scale datasets. We collected videos from open source video datasets such as Li et al. (2024) and Wang et al. (2024), followed by an initial coarse filtering process that detected the presence of human-related descriptions in video captions. It is worth noting that the captions provided by these datasets are inherently coarse-grained and often fail to capture the nuanced, dynamic activities performed by characters (e.g., complex gestures, interactions, or context-specific behaviors). To address this limitation, we developed a specialized captioning pipeline designed to focus on human motion patterns, which will be elaborated in subsequent subsections.

Manual curation of high-quality samples. Complementing the above approach, we manually selected videos containing intentional and complex human activities (e.g. speaking, singing, dancing) from public accessible sources. This dual methodology yielded an initial video pool comprising millions of human-centric video samples, forming the foundation for our dataset.

Pose Tracking and Fine-grained Filtering. From the initial human-centric video pool, the 2D pose of each character is tracked via VitPose Xu et al. (2022) and converted to DWPose Yang et al. (2023). This pose information serves two critical functions: (1) As a multi-modal control signal: The tracked

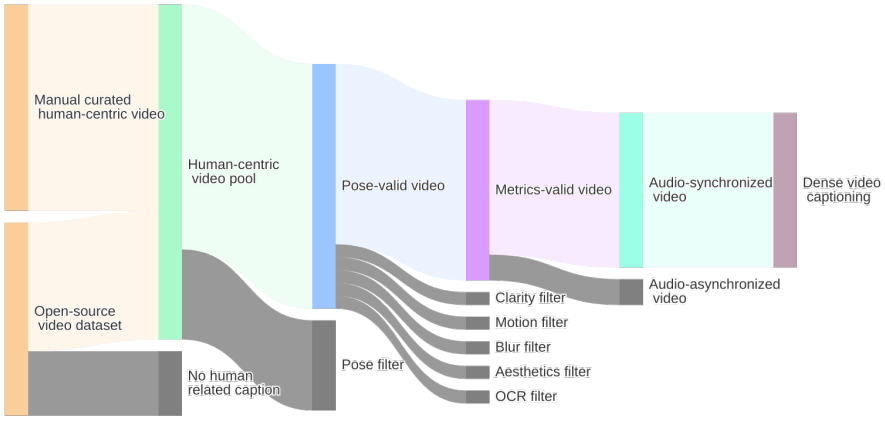

Figure 1: Overview of our hierarchical human-centric video filtering pipeline.

pose is integrated as an optional multi-modal control signals for our human-centric video generation model, enabling precise temporal alignment with human actions. (2) For dataset refinement: The pose data is further leveraged to implement a fine-grained screening process. Specifically, we filtered out videos where characters occupy only a negligible portion in either temporal or spatial dimensions. Additionally, to ensure the model can learn audio-driven facial expressions from the given audio signals, we retained only videos containing consistent and visible human faces throughout the sequence. Complementing the pose-based screening, we employed pre-trained video quality assessment models to evaluate motion extent, aesthetic appeal, and visual clarity. Videos were subsequently filtered based on these quantitative metrics to maintain high data quality. Furthermore, to address audio-visual alignment challenges, we utilized Light-ASD Liao et al. (2023) to detect and exclude videos where (1) the audio is not synchronized with the active speaker, or (2) no active speaker exists in the scene.

Video Quality. To comprehensively evaluate video quality from multiple perspectives, we employ the following five metrics: (1) Clarity Assessment: We utilize the Dover metric Wu et al. (2023) to quantify video clarity, which measures the perceptual sharpness of visual content. (2) Motion Stability Analysis: To evaluate temporal coherence, we predict optical flow using the UniMatch framework Xu et al. (2023) and calculate a motion score. This helps identify and filter videos with excessive subject/background movement that could compromise visual quality. (3) Facial/Hand Sharpness Verification: A Laplacian operator is applied specifically to human faces and hands within the video frames. This technique enables the detection and exclusion of videos containing blurred facial features or hand regions. (4) Aesthetic Quality Evaluation: We incorporate an improved aesthetic predictor Schuhmann (2022) to assess visual appeal based on human aesthetic preferences, ensuring the output meets subjective quality standards. (5) Subtitle Occlusion Detection: An OCR-based detector is applied to identify and exclude cases where subtitles might occlude faces or hands in video.

Dense Video Caption. A detailed and accurate video caption facilitates the alignment of the generation model with the input prompt. We used QwenVL2.5-72B Bai et al. (2025) to generate captions for the videos, instructing the model to describe the following key aspects in details: (1) Camera angles, such as straight-on, overhead, low-angle, wide shot, medium shot, and close-up; (2) Physical appearance features (e.g., clothing and accessories) and actions, broken down into specific movements of the subject; (3) Main features of the background environment, including architectural style, color schemes, and greenery, among others. At the same time, we required the model to avoid subjective evaluations and emotional interpretations, which are often trivial to generating the expected video content

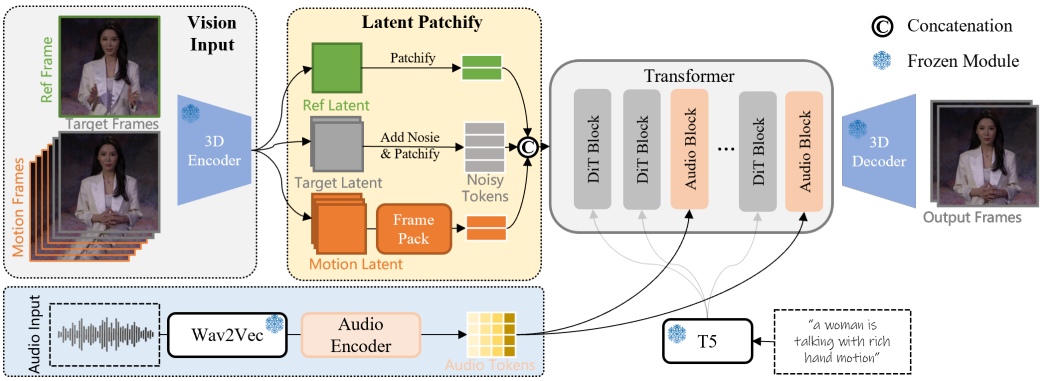

Figure 2: Overview of our pipeline.

3 Model Architecture

Given a single reference image, an input audio and a prompt to describe the video content, we could generate the video synchronized with the audio while preserving the content in the reference image (not start from the image). As shown in Fig 2, our work is fed with multi-frame noise latent input, and tries to denoise them to the consecutive video frames during each timestep.

During the training, the RGB target frames $X \in \mathbb{R}^{F\times H\times W\times 3}$ are encoded by 3D VAE into latent presentation $x_0 \in \mathbb{R}^{f\times h\times w\times c}$, assigning a continuous time step $t \in [0, 1]$, the noise ϵ are added to x_0 to get noisy latent x_t according to Flow Matching introduced by Lipman et al. (2023):

$$x_t = t\epsilon + (1-t)x_0$$

Input the noisy representation x_t, the target of the model is to predict the velocity $\frac{dx}{dt} = \epsilon - x_0$. During the inference, the model recoveres the noisy input x_t into x_0 under the condition of the reference frame, motion frames, audio input and prompt.

The reference image, the target frames and the motion frames following Tian et al. (2024) are fed into 3D VAE to down-sample the video spatially and temporally, getting the latent representation of the frames. All latent frames are then patchified and flattened, they are concatenated to be a sequence of visual tokens. The motion frames are optional, they provide the condition of the previous information, making the generated clips continuous. In order to generate long-term consistent video frames, it is necessary to obtain more historical information, since directly flatten the motion latent token could introduce more computational load. The motion latent is further compressed by Frame Pack module introduced by Zhang & Agrawala (2025), which compress the earlier frames in higher compressibility.

As illustrated in Figure 3, the raw audio waveform is first encoded using Wav2Vec by Schneider et al. (2019). To comprehensively capture the audio features, we adopt the weighted average layer proposed by Tian et al. (2024), which combines features from different layers through learnable weights. This approach effectively integrates shallow-level rhythmic and emotional cues with deep-level lexical content features extracted by Wav2Vec, thereby enhancing synchronization with complex audio signals such as singing or expressive speech. The resulting frame-wise audio features are then compressed along the temporal dimension using multiple causal 1D convolutional modules. This process generates audio features of the i th latent frame $a_i \in \mathbb{R}^{f\times t\times c}$ that are temporally aligned with the video latent frames, where t denotes the number of audio tokens per latent frame.

The latent audio features a are passed into each Audio Block, where the noisy latent tokens $x_t \in \mathbb{R}^{(f'\times h\times w)\times c}$ are divided into segments $\sum_i^{f'} x_{ti} \in \mathbb{R}^{(h\times w)\times c}$ along the temporal dimension. To reduce computational overhead, attention is calculated between a_i and x_{ti}, rather than performing full 3D attention between visual tokens and audio tokens. This approach ensures that the audio features and visual tokens are naturally synchronized.

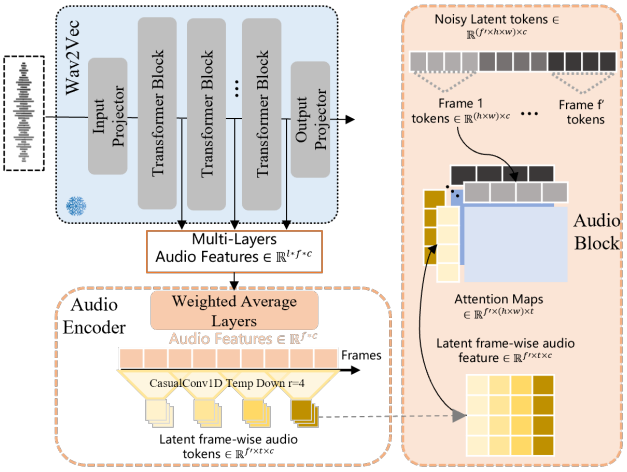

Figure 3: The pipeline of the audio injection.

4 IMPLEMENTATION

To train our Audio-to-Video model, we adopted a hybrid-parallel training scheme which combines FSDP and context parallelism, enabling large-scale, full-parameter model training. To support different resolutions, we support the training of variable-length video data. Our model is trained based on a pre-trained Wan model and is designed with a three-stage training process, including: audio encoder training, training on speech videos, training on film and television + speech videos, and finally, high-quality SFT (Supervised Fine-Tuning) stage training.

4.1 PARALLEL STRATEGY

To efficiently train our large model, a hybrid parallel training strategy is employed. This involves combining Fully Sharded Data Parallelism (FSDP) Zhao et al. (2023) with Context Parallelism. Initially, FSDP is leveraged to shard the model's parameters across 8 GPU cards within a single node, enabling the training of our Wan-S2V-14B model while utilizing 80GB of memory per GPU.

Subsequently, for parallel computing, we implement a Context Parallelism scheme, combining RingAttention and Ulysses similar to Fang & Zhao (2024). This integrated approach, executed on 8 GPUs within a single node, allows us to achieve near-linear speedup, significantly reducing the single training iteration time from $\sim$100 seconds to $\sim$12 seconds. This robust setup ultimately supports the training of models exceeding 16B parameters, including our audio encoder and cross-attention components, enabling high-resolution video training up to 48 frames at 1024×768 resolution (Height $\times$ Width) on 8 GPUs.

To accommodate diverse output resolutions and optimize training, a variable-length resolution training method is implemented. This method uses the token count, determined after the patchify operation, as a key metric. A maximum allowable token limit, M, is established. For videos exceeding this limit, resolution resizing or cropping is applied to reduce the token count to M or below. Videos with token counts already below M are used directly for model training without any modifications.

5 EXPERIMENTS

Following the data construction pipeline detailed in Section 3, we meticulously filtered data from the OpenHumanViD Li et al. (2024) dataset and integrated it with our self-constructed internal talking head dataset to form our comprehensive training set.

We constructed the audio-driven human video generation model on Wan-14B referred to as Wan-S2V-14B.

In comprehensive comparisons against existing state-of-the-art audio-driven video generation models, both quantitative metrics and visual results consistently demonstrate that our method surpasses current approaches in terms of expressiveness and the realism of generated content.

5.1 Qualitative Evaluation

Comparison with SOTA

A comparative study was conducted between our method and two existing DiT-based audio-driven video generation models, Ominihuman proposed by Lin et al. (2025) and Hunyuan-Avatar proposed by Chen et al. (2025), revealing the superior capabilities of our approach. Figure 4 illustrates these findings: Hunyuan-Avatar struggles with facial distortion and identity shifts during large-scale movements, while our model excels at maintaining identity consistency even amidst highly dynamic motion. Additionally, Ominihuman's generated results are characterized by very small motion amplitudes, often closely resembling the reference image's static pose. Our model, conversely, is capable of generating a significantly wider range of motion, thus offering enhanced diversity in output.

Consistency of Long Term Generation

Compared to previous methods that typically generate short, isolated video clips focused on solo speaking scenarios, film-grade video generation demands long-term consistency across multiple generated clips, e.g motion, camera movement and identity preservation. Our method utilizes FramePack to encode more motion frames, enabling the model to capture long-term temporal dependencies and, intuitively, achieve better preservation of coherent temporal information.

As shown in 5, when generating a scene in which the target is required to maintain consistent motion (e.g., a train moving in a coherent direction), OmniHuman fails to preserve the motion trend across multiple clips, while our method successfully maintains consistency in both the direction and speed of the train.

When continuing to generate a new video clip following previously generated ones, the prior clips are used as motion frames. By utilizing FramePack to encode a larger number of motion frames, our method not only preserves the overall motion trend but also helps maintain element identity across clips. For instance, as shown in 6, the generated character picks up a piece of paper that visually matches the one from the previous clip. In contrast, without FramePack, the appearance of the same object may drift significantly.

5.2 Quantitative Evaluation

We conduct quantitative comparisons on the EMTD dataset proposed by Meng et al. (2024),which primarily consists of solo-talking videos, evaluating several open-source audio-animation methods. This includes EchoMimicV2, developed by Meng et al. (2024), and MimicMotion from Zhang et al. (2024). Both of these approaches rely on pre-extracted pose sequences to animate images. Additionally, we compare our work with EMO2, introduced by Tian et al. (2025a), which employs a two-stage process: generating partial hand motion from audio and subsequently animating the character using both the audio and the generated motion. We also include recent audio-driven DIT-based methods in our comparisons, such as FantasyTalking Wang et al. (2025) and Hunyuan-Avatar.

To demonstrate the superiority of our proposed method, we evaluate the models using several metrics. We employ Fréchet Inception Distance (FID) Heusel et al. (2017), SSIM Wang et al. (2004), and PSNR Horé & Ziou (2010) to assess the quality of the generated frames. Fréchet Video Distance (FVD) Unterthiner et al. (2019) is used to gauge the overall coherence of the generated videos. To evaluate identity consistency, we calculate the cosine similarity (CSIM) between the facial features of the reference image and the generated video frames. We also utilize Sync-C, as proposed by Chung & Zisserman (2017), to assess the synchronization quality between lip movements and audio signals. Furthermore, we measure Hand Keypoint Confidence (HKC) to evaluate the quality of hand representation in generated frames, while Hand Keypoint Variance (HKV) serves as an indicator of the richness of hand motion. Additionally, EFID proposed by Tian et al. (2025b) is adopted to quantitatively assess the divergence in expressions between the synthesized videos and those in the ground truth dataset.

Figure 4: Qualitative comparison of generated human videos. The leftmost column displays the reference image. Hunyuan-Avatar (top row) often suffers from facial distortions and inconsistent identity during large movements. Ominihuman (middle row) typically generates results with a limited range of motion, largely adhering to the pose of the reference image. In contrast, our method (bottom row) achieves superior performance in both motion dynamics and identity consistency.

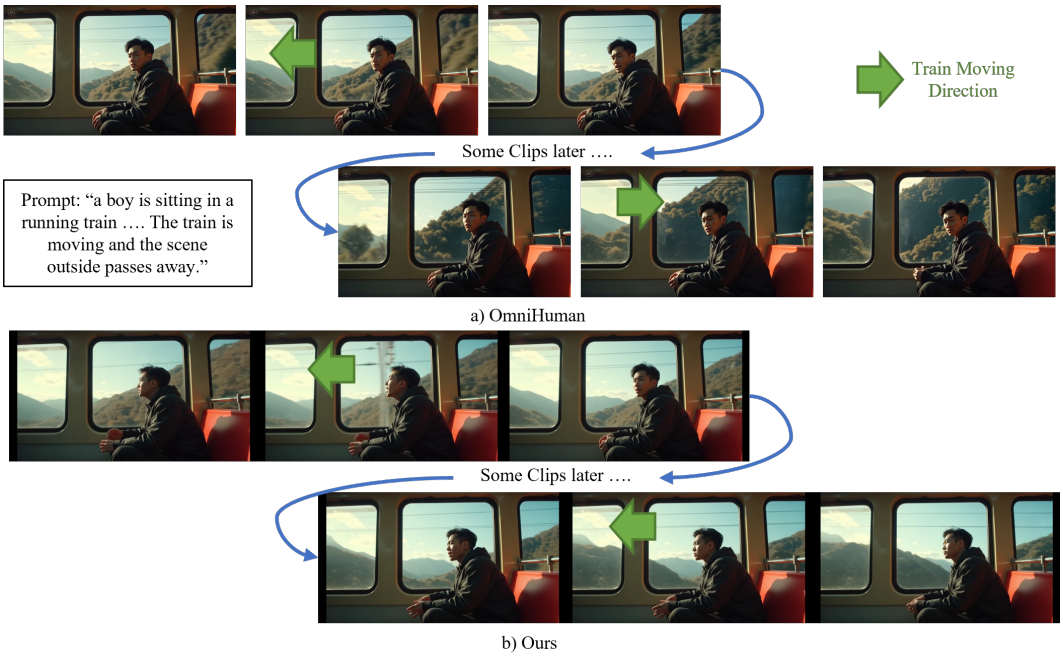

Figure 5: Qualitative comparison of motion preservation performance between our method and OmniHuman.

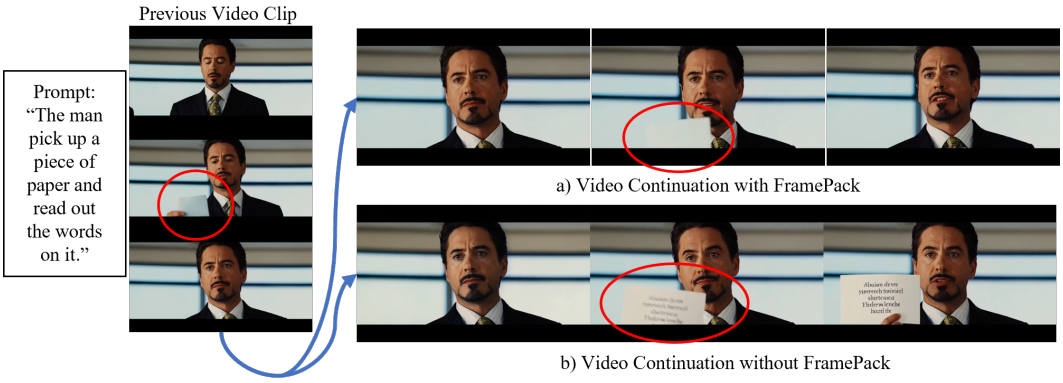

Figure 6: Maintaining item identity across consecutive video clips.

Table 1: Quantitative comparisons with SOTA.

Method	FID↓	FVD↓	SSIM↑	PSNR↑	Sync-C↑	EFID↓	HKC↑	HKV↑	CSIM↑
EchoMimicV2	33.42	217.71	0.662	18.17	4.44	1.052	0.425	0.150	0.519
MimicMotion	25.38	248.95	0.585	17.15	2.68	0.617	0.356	0.169	0.608
EMO2	27.28	129.41	0.662	17.75	4.58	0.218	0.553	0.198	0.650
FantasyTalking	22.60	178.12	0.703	19.63	3.00	0.366	0.281	0.087	0.626
HY-Avatar	18.07	145.77	0.670	18.16	4.71	0.7082	0.379	0.145	0.583
Ours	15.66	129.57	0.734	20.49	4.51	0.283	0.435	0.142	0.677

As illustrated in 1, our method surpasses the others in terms of frame quality, as indicated by improved image metrics (FID, SSIM, PSNR). Additionally, it demonstrates a clear advantage in video quality assessment, with a lower FVD score. In terms of detail generation, our approach produces clearer and more accurate hand shapes, as reflected by the higher HKC score. Furthermore, it generates more vivid and diverse hand motions, indicated by a higher HKV value. It is worth noting that EMO2 achieves highest HKC and HKV scores. This can be attributed to the fact that EMO2 generates frames conditioned on pre-generated motion sequences, allowing for better control over hand motion diversity. Moreover, the use of MANO contributes to its superior performance in HKC compared to other methods. On the other hand, HY-Avatar tends to produce characters with "poker-face" expressions, which results in a higher EFID compared to other methods.

6 Conclusion

This paper presented significant advancements in audio-driven human video generation, specifically addressing the complexities of film and television scenarios. We demonstrated the crucial synergy between text for global motion control and audio for fine-grained character expressions, leading to more expressive and consistent video generation. Our comprehensive approach, from data to training and optimized inference, aims to make high-quality audio-driven video synthesis more accessible and practical. Despite this progress, truly complex film and television challenges, such as nuanced multi-person interactions and precise camera control driven solely by audio, remain formidable. Wan-S2V is the first in our Vida research series. We envision this series, including future work on advanced character control and dynamic dancing generation, will foster continued research and development, pushing the boundaries of human-centric video synthesis.

7 Contributors

All contributors are listed in alphabetical order by their last names.

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, Lian Zhuo

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters, 2025. URL https://arxiv.org/pdf/2505.20156.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pp. 251–263. Springer, 2017.

Jiarui Fang and Shangchun Zhao. Usp: A unified sequence parallelism approach for long context generative ai, 2024. URL https://arxiv.org/abs/2405.07719.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pp. 2366–2369, 2010. doi: 10.1109/ICPR.2010.579.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Daquan Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, and Siyu Zhu. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation, 2024.

Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, and Liangyin Chen. A light weight model for active speaker detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22932–22941, June 2023.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models, 2025. URL https://arxiv.org/abs/2502.01061.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation, 2024. URL https://arxiv.org/abs/2411.10061.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. pp. 3465–3469, 09 2019. doi: 10.21437/Interspeech.2019-1873.

Christoph Schuhmann. improved-aesthetic-predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive - generating expressive portrait videos with audio2video diffusion model under weak conditions, 2024.

Linrui Tian, Siqi Hu, Qi Wang, Bang Zhang, and Liefeng Bo. Emo2: End-effector guided audio-driven avatar video generation, 2025a. URL https://arxiv.org/abs/2501.10687.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pp. 244–260. Springer, 2025b.

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

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Mengchao Wang, Qiang Wang, Fan Jiang, Yaqi Fan, Yunpeng Zhang, Yonggang Qi, Kun Zhao, and Mu Xu. Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. *ArXiv*, abs/2504.04842, 2025. URL https://api.semanticscholar.org/CorpusID:277621659.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, and Di Zhang. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content, 2024. URL https://arxiv.org/abs/2410.08260.

Zhou 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. doi: 10.1109/TIP.2003.819861.

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 *International Conference on Computer Vision (ICCV)*, 2023.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation, 2022. URL https://arxiv.org/abs/2204.12484.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4210–4220, 2023.

Lvmin Zhang and Maneesh Agrawala. Packing input frame contexts in next-frame prediction models for video generation. *Arxiv*, 2025.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. Pytorch fsdp: Experiences on scaling fully sharded data parallel, 2023. URL https://arxiv.org/abs/2304.11277.