StableAvatar: Ultra-Long Audio-Driven Avatar Video Generation

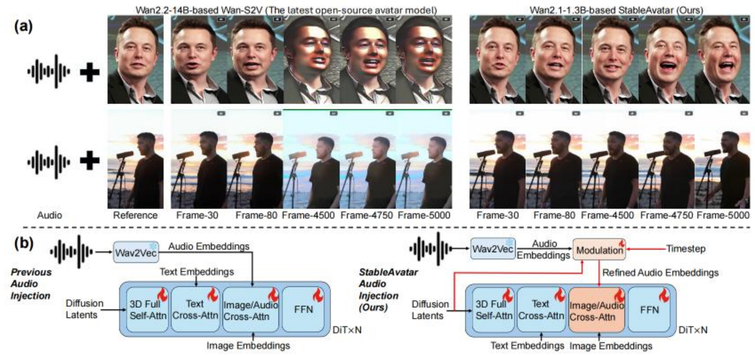
Shuyuan Tu[1] · Yueming Pan[3] · Yinming Huang[1] · Xintong Han[4] · Zhen Xing[1] · Qi Dai[2] · Chong Luo[2] · Zuxuan Wu[1,*] · Yu-Gang Jiang[1]

[1]Fudan University [2]Microsoft Research Asia [3]Xi'an Jiaotong University [4]Tencent Inc.

CODE PROJECT PAGE

NEURAL INFORMATION PROCESSING SYSTEMS

Overview & Methods

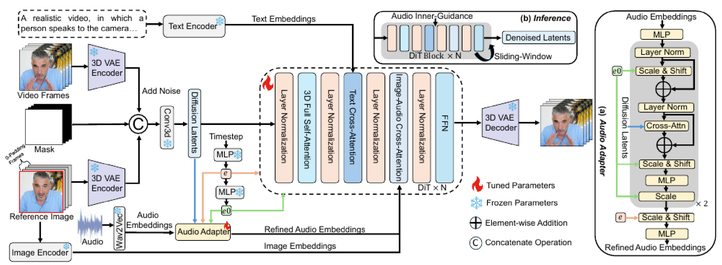


Motivation

Long generation: identity, visual quality and lip synchronization degrade as clip-level errors accumulate.
Audio mismatch: off-the-shelf embeddings do not follow the diffusion timestep or the pretrained video prior.

Contributions

(1) We propose a novel Timestep-aware Audio Adapter to force the diffusion model to capture joint audio-latent features, thereby significantly reducing latent distribution error accumulation during audio injection, making StableAvatar the first video diffusion transformer that generates ultra-long smooth audio-driven avatar video.
(2) A novel Audio Native Guidance Mechanism replaces the conventional CFG to further enhance audio synchronization. At the same time, a dynamic weighted sliding-window denoising strategy improves the smoothness of synthesized long videos.
(3) Experiments on benchmarks show the superiority of StableAvatar over the SOTA. Notably, based on a Wan2.1-1.3B model, we surpass previous Wan2.1-14B-based models in video quality for long-video generation.



Timestep-aware Audio Adapter

At each diffusion step, the adapter maps audio features into the current latent space before cross-modal attention.

Audio Native Guidance

$$(1+\alpha+\beta)\boldsymbol{D}([\boldsymbol{z}_t,\bar{\boldsymbol{a}}_t],\boldsymbol{y},\boldsymbol{I},\boldsymbol{A};\theta) - \alpha\boldsymbol{D}([\boldsymbol{z}_t,\bar{\boldsymbol{a}}_t],\boldsymbol{y},\boldsymbol{I},\emptyset;\theta) - \beta\boldsymbol{D}([\boldsymbol{z}_t,\emptyset],\boldsymbol{y},\boldsymbol{I},\boldsymbol{A};\theta),$$

ANG guides denoising with the network's own audio representation, which improves motion and lip synchronization.

Dynamic Weighted Sliding Window

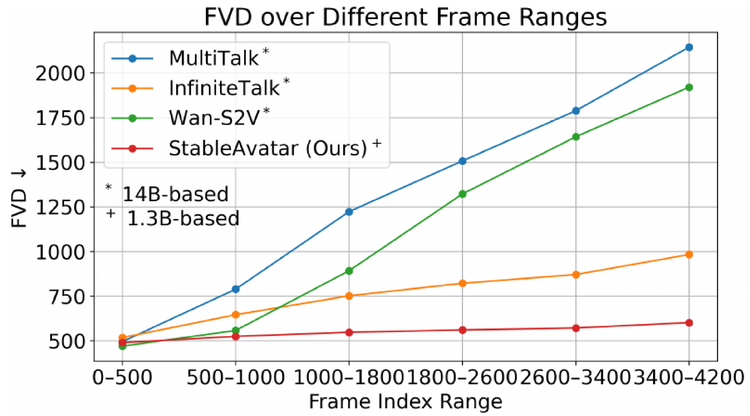
Position-aware blending gives central frames more weight and suppresses unstable boundary predictions.

Experiment Results

Table 1: Quantitative comparisons on HDTF/AVSpeech/Long100. In the table elements a / b / c, a, b, and c refer to the result on the HDTF, AVSpeech, and Long100, respectively. The average video duration of HDTF and AVSpeech is 10 seconds, while Long100 is 3 minutes.

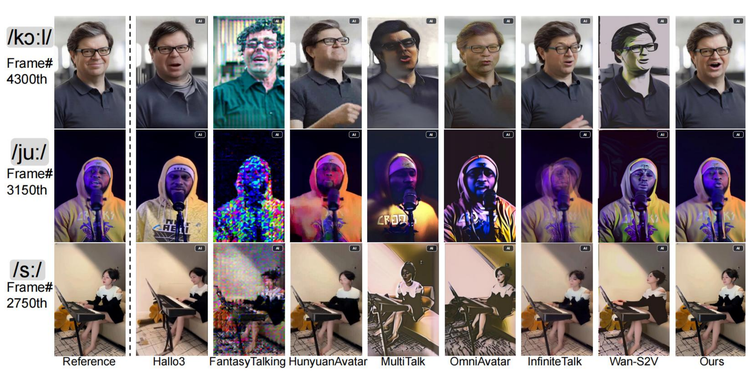
Model	FID↓	FVD↓	CSIM↑	Sync-C↑	Sync-D↓	IQA↑	ASE↑
SadTalker [87]	53.12/120.57/194.88	567/1468/2261	0.821/0.802/0.423	7.25/4.13/3.21	9.86/9.72/11.16	3.12/2.40/2.31	2.14/1.43/1.38
Sonic [30]	62.17/187.42/278.40	552/2051/2877	0.843/0.825/0.451	8.16/6.04/4.03	7.65/8.94/10.23	3.28/2.24/2.08	2.05/1.36/1.13
Hallo3 [11]	44.31/98.14/170.44	438/987/1724	0.845/0.834/0.462	6.58/5.16/4.42	8.64/9.62/9.92	3.50/3.38/2.35	2.11/1.96/1.36
FantasyTalking [77]	46.74/80.01/175.78	479/823/1789	0.863/0.853/0.468	3.42/2.98/1.92	12.15/11.42/11.78	3.55/3.24/2.42	2.28/1.89/1.48
HunyuanAvatar [6]	52.16/77.53/172.94	625/868/1743	0.866/0.859/0.472	7.20/6.74/4.34	8.39/8.28/10.07	3.56/3.63/2.46	2.25/2.21/1.52
MultiTalk [33]	46.94/75.67/175.52	446/804/1768	0.868/0.861/0.465	7.53/4.88/4.12	8.02/9.59/10.18	3.54/3.72/2.42	2.15/2.25/1.45
OmniAvatar [15]	41.79/72.56/168.49	424/744/1621	0.862/0.857/0.471	7.50/6.78/4.45	8.26/8.05/9.62	3.55/3.74/2.51	2.27/2.29/1.56
EchomimicV3 [45]	60.34/106.31/172.18	590/1108/1852	0.852/0.840/0.458	5.48/4.62/3.71	9.33/9.78/10.24	3.66/3.04/2.38	2.32/1.85/1.39
InfiniteTalk [84]	40.12/69.48/68.95	390/652/569	0.869/0.865/0.755	7.97/7.23/7.92	7.34/8.15/7.05	3.76/3.77/3.65	2.38/2.28/2.26
Wan-S2V [16]	**35.87/66.43**/164.42	**368/625**/1685	0.872/0.869/0.477	8.12/7.48/4.88	6.98/7.92/9.68	3.83/**3.81**/2.47	2.42/2.30/1.53
Ours	38.14/68.12/**57.18**	375/640/**504**	**0.875/0.872/0.849**	**8.15/7.56/8.24**	**6.94/7.85/6.79**	**3.90**/3.79/**3.84**	**2.46/2.32/2.39**

Quality drift across frame ranges



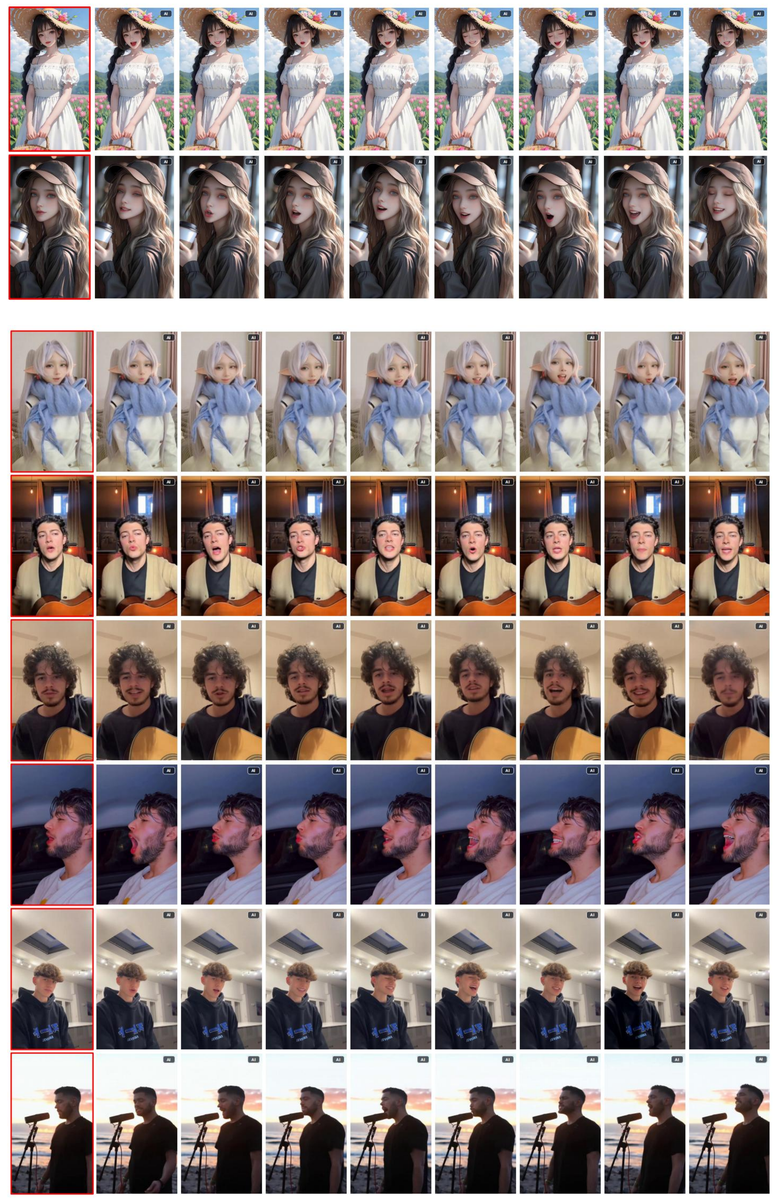
StableAvatar accurately animates the reference with natural lip synchronization driven by audio while preserving the overall consistency, even in the ultra-long video generation scenario (3500+ frames in a single pass).

Representative cases



More Cases

Long-video generations



* Corresponding author | Synthesized frames are watermarked to mitigate deepfake impersonation risks | NeurIPS 2026