TalkVerse: Democratizing Minute-Long Audio-Driven Video Generation

Supplementary Material

1. More Details about Experiments

1.1. Training Details

We finetune the Wan S2V 5B model and newly added audio cross-attention using AdamW (10^{-5} learning rate) with selective unfreezing of the time embedding, projection, head, and optional transformer blocks. LoRA parameters in DiT are scaled by $10\times$ to be 10^{-4}. Training uses a batch size of 1 per GPU with gradient accumulation to be total batch size as 256 and gradient clipping (norm 1.0). To perform CFG, we apply 10% dropout to text, image, and audio conditions. When framepack module is enabled, we jointly encodes motion and target frames to VAE space for training to maintain temporal continuity in long-video generation. Otherwise, the motion context frames in VAE space as all-zero tensors.

1.2. Spatial Loss Weighting

We use a ROI loss up weights for reconstruction error in face and body regions by employing DWPose [1] to estimate body and face bounding boxes and keypoints which are normalized and rasterized into binary masks aligned with the resolution in VAE space, and then construct a spatial loss weighting over the diffusion denoising objective. Specifically, let L_{full}, L_{body}, and L_{face} denote the per-pixel losses on the full frame, the body bounding box, and the face bounding box, respectively. The final training loss is

$$L = \frac{1 \cdot L_{\text{full}} + 1 \cdot L_{\text{body}} + 1 \cdot L_{\text{face}}}{Z}, \tag{1}$$

where Z is a normalization factor (e.g., the sum of the weights), so that regions inside the body and face boxes receive proportionally higher importance than the background.

1.3. Data Preprocessing Pipeline

We construct a video–audio–text dataset by unifying metadata from multiple sources, ensuring all clips contain valid 25 fps video and audio streams. To control resolution, we adopt a resolution-bin policy (e.g., 720p) where each clip is resized and center-cropped to the closest target bucket. We sample a continuous sequence of frames for denoising, optionally prepended by motion context frames, and extract a conditional reference image from a nearby window.

Audio is resampled to 16 kHz and strictly aligned to the video segment, padded if necessary. Finally, we load cached text embeddings (pre-extracted by T5 text encoder) and a null embedding for classifier-free guidance, yielding a comprehensive set of synchronized inputs for training.

Table 1. Hyper-parameters for our 5B model.

Hyper-parameter	Value
Language Encoder	T5-XXL
VAE	Wan2.2-VAE
VAE stride	$(4, 16, 16)$
Audio encoder	wav2vec2-large-xlsr-53-english
Patch size	$(1, 2, 2)$
Model dimension	3072
FFN dimension	14336
Frequency embedding dimension	256
Attention heads	24
Transformer layers	30
QK norm / cross-attn norm	True / True
Audio dimension	1024
Audio injection layers	$\{0, 3, 6, 9, 12, 15, 18, 21, 24, 27, 29\}$
Motion frames	73
Latent dimension	48
v-scale	1.0
Sampling frame rate	25 fps
Sampling frame count	120
Diffusion shifting	5
Default sampling steps	50

1.4. Inference Process

Our inference pipeline supports both image-to-video generation and video-to-video "dubbing" across multiple resolutions. We encode the text prompt (T5), audio (wav2vec2), and reference image (VAE) to condition the diffusion model. For dubbing, instead of initializing from pure noise, we encode the input video into latents, add noise to a target timestep t_{dub}, and denoise from that state to preserve coarse structure while synchronizing lip motion. The latent sequence is iteratively denoised using a flow-matching solver (e.g., UniPC, 50 steps) and decoded by the VAE. For CFG scale, we find that 3 to 6 will leads to good results, which is a trade-off between (1) visual quality and motion strength, and (2) lip-sync quality. The larger CFG weights, the better lip-sync quality but worse visual quality and motion strength.

1.5. Hyper-parameter Settings for Wan S2V 5B

We summarize the main configuration used in our experiments in Table 1.

References

[1] 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*, pages 4210–4220, 2023. 1