DUET: AUDIO-DRIVEN HUMAN VIDEO SYNTHESIS WITH VOICE-FACE DUAL HARMONIZATION

Anonymous authors
Paper under double-blind review



Figure 1: Harmonized audio-driven audio-video generation with speaker-ID binding. Given driving audio clips and a conditional image, our method synthesizes a synchronized video while adapting the generated timbre toward the vocal identity of the corresponding on-screen speakers.

ABSTRACT

Existing audio-driven human video generation predominantly focuses on lip synchronization, yet often overlooks the timbre mismatch between the driving voice and the depicted speaker. We propose DUET, a voice-face dual-conditioned talking video generation model that animates the face with lips synchronized with the driving audio while adaptively updating the audio timbre to match the speaker appearances. We formulate this task within a joint audio-video generation framework: the driving utterance provides linguistic and temporal cues, while the conditional image determines the vocal identity of the generated speech. We incorporate a Whisper-based audio injection branch into a pretrained joint audio-video generative model and train it on timbre-perturbed audiovisual pairs to facilitate both lip synchronization and timbre reasoning through cross-modal learning. For multi-person conversational scenes, we introduce a spatio-temporal masking mechanism that associates each audio condition with its designated face region and speaking interval. On the speaker-identity-disjoint EMTD benchmark, our method obtains the best or tied-best scores on the evaluated voice-conversion metrics. On EMTD and InterActEyes, it maintains competitive visual quality and lip synchronization. In a perceptual comparison with a DreamVoice+MultiTalk cascade on EMTD, participants prefer our method in 64.29% of responses for timbre–identity consistency and 53.97% for overall audiovisual coherence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Audio-driven human animation (Lin et al., 2025; Wang et al., 2025b;c; Gao et al., 2025) seeks to synthesize realistic videos where facial expressions, lip articulations, and bodily gestures are temporally and semantically anchored to a driving acoustic signal. Video generative foundation models have improved applications in digital humans, automated dubbing, and personalized media. Existing methods, however, primarily optimize lip synchronization and often inherit the driving speaker's timbre even when it conflicts with the visible identity. The mismatch is especially apparent when the apparent age or vocal presentation of a character differs substantially from that of the driving speaker, as illustrated in Fig. 1.

The driving audio should not be treated as a single undifferentiated condition. Natural speech contains linguistic content, rhythm, prosody, and speaker timbre, which play different roles. Linguistic and time-varying cues should control speech motion and facial dynamics. Source-speaker timbre should not be copied when it conflicts with the visible identity. The conditional image instead supplies the vocal identity of the on-screen speaker. In short, the driving audio tells the model what is said and how it is spoken, but not who should be heard.

A straightforward alternative is to first perform text- or face-guided voice conversion (Hai et al., 2024; Sheng et al., 2023) and then drive a talking-video generator (Kong et al., 2025) with the converted speech. However, standalone voice conversion may be suboptimal, as predicting a target-consistent voice in the wild relies on rich cross-modal priors. We instead build on a pretrained audiovisual generative model to better exploit such priors. Moreover, cascaded pipelines can accumulate errors, since imperfections in the converted speech propagate to downstream video generation. In contrast, we jointly model audio and video, allowing facial dynamics, lip motion, and timbre to interact within a unified cross-modal process. We compare this joint formulation with a DreamVoice+MultiTalk cascade in Sec. 4. The portrait not only provides the identity to animate but also conditions the audio synthesized jointly with the video.

This setting poses three challenges. First, speech content and speaker timbre are entangled in natural audio, making it difficult to retain useful driving cues without copying the source voice (Liu, 2024; Li et al., 2023; Qian et al., 2019). Second, no target voice recording is provided at inference: the model must infer vocal identity from visual appearance rather than cloning a reference voice sample. We evaluate this inferred identity by comparing the generated speech with the original paired recording of the on-screen speaker. Third, in two-person conversations, each audio stream must remain associated with its designated face and speaking interval.

These challenges require separating transferable linguistic and temporal cues from speaker identity, which should instead arise from the visual condition. They also call for evaluating generated speech as part of the audiovisual output, rather than treating audio only as an animation input. We therefore frame audio-driven animation as a harmonization problem across motion, content, and perceived identity.

We build DUET on a pretrained joint audio-video generative model. The conditional image serves as the initial frame, while a Whisper-based branch injects linguistic and temporal cues from the driving audio. Training on timbre-perturbed audiovisual pairs discourages copying the driving-speaker timbre and encourages the generated voice to approach the original paired voice of the visible speaker.

We further introduce a spatio-temporal masking mechanism for two-person conversational scenes. The mechanism associates each audio condition with its designated facial region and speaking interval. Experiments and a perceptual study show closer agreement with the paired original voice while maintaining competitive lip synchronization and visual quality.

Our contributions are summarized as follows:

1. We identify a common limitation of existing audio-driven video generation formulation and pioneer a practically favorable voice-face dual harmonization paradigm that advances this technique toward real-world application demands.
2. We propose DUET, a voice-face dual-conditioned video generation model that effectively achieves audio-visual synchronization, speaker-identity coherent timbre, and high conformity to the linguistic and emotional cues of the driving speech.

3. We design a spatio-temporal voice-face binding mechanism for multi-person conversations and demonstrate our superiority over strong baselines across multiple evaluation settings.

2 Related Work

2.1 Voice Conversion

Voice conversion (VC) aims to modify the speaker identity or vocal timbre of a source utterance while preserving its linguistic content and prosodic information. Conventional VC methods typically obtain the target timbre from a specified target speech recording or speaker embedding. Representative neural methods (Kaneko & Kameoka, 2018; Kameoka et al., 2018; Qian et al., 2019) improve the flexibility of VC through adversarial learning or representation disentanglement. More recent VC methods (Li et al., 2023; Liu, 2024; Popov et al., 2022; Liu et al., 2021; Cai et al., 2025) extract linguistic content from source speech using self-supervised speech models and synthesize the converted waveform with high-fidelity neural vocoders or diffusion-based generators.

Audio-reference-based VC methods require target speech or speaker embeddings, which has motivated work that predicts target timbre from text or visual cues. Text-guided VC methods (Yao et al., 2024; Hai et al., 2024) convert voices from natural-language descriptions, while face-based VC methods (Lu et al., 2021; Weng et al., 2023; Sheng et al., 2023; Rong & Liu, 2025) estimate vocal characteristics from facial images. These studies show that timbre can be obtained without a target-audio recording, but they operate in the audio domain and do not jointly synthesize a synchronized video.

2.2 Cross-Modal Face–Voice Generation

Face and voice carry correlated but non-identical information. Speech2Face (Oh et al., 2019) reconstructs facial appearance from speech, illustrating both the learnable cross-modal correlation and its ambiguity. Faces that Speak (Jang et al., 2024) jointly synthesizes speech and a talking face from text and a portrait, while FaceSpeak (Zhang et al., 2025b) generates portrait-aligned expressive speech from text. These works show that visual context can guide vocal attributes. Our setting differs in input and objective: rather than synthesizing speech content from text or cloning a target voice recording, we take an existing driving utterance, retain its linguistic and temporal cues, replace its source-speaker timbre, and jointly generate audio and video. At test time we measure closeness to the original paired recording of the on-screen speaker. We further consider explicit voice-face assignment in two-person conversations.

2.3 Audio-driven Human Animation

Given speech signals and reference identity images, audio-driven human animation synthesizes temporally coherent human videos. Earlier approaches (Prajwal et al., 2020; Zhou et al., 2020; 2021; Wang et al., 2021; Richard et al., 2021; Fan et al., 2022; Zhang et al., 2023b) often rely on landmarks, 3D facial parameters, meshes, or disentangled audio-visual embeddings. Diffusion-based methods (Shen et al., 2023; Stan et al., 2023; Sun et al., 2024; Tian et al., 2024) treat audio-driven facial animation as a conditional generative process. With video foundation models (Yang et al., 2025; Wan et al., 2025; Kong et al., 2024), recent works (Wang et al., 2025b; Lin et al., 2025; Gan et al., 2025; Tu et al., 2025; Gao et al., 2025) leverage video generative priors for portrait or human animation. Multi-person talking-video methods (Kong et al., 2025; Huang et al., 2025; Zhong et al., 2025; Wang et al., 2025c) bind multiple audio streams to corresponding identities, but they take the driving audio as given and do not regenerate a visually consistent timbre.

2.4 Audio-Video Joint Generation

Audio-video joint generation synthesizes synchronized visual and acoustic content end-to-end, avoiding the post-hoc synchronization issues of cascaded pipelines. Proprietary systems such as Veo 3 (Google DeepMind, 2025) and Sora 2 (OpenAI, 2025) generate native audio-video from text. In the open-source community, UniVerse-1 (Wang et al., 2025a) stitches pretrained audio and video experts, while Ovi (Low et al., 2025), Harmony (Hu et al., 2025), and UniAVGen (Zhang

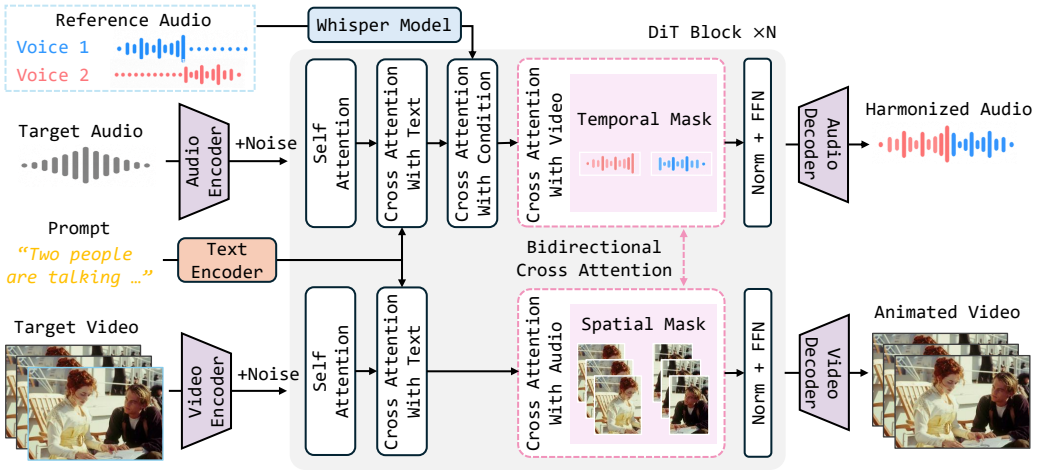

Figure 2: Overview of the proposed framework. Two parallel denoising branches jointly generate audio and video. Whisper-encoded driving audio is injected through cross-attention, while spatio-temporal masks associate each generated-audio interval with its designated visual region in two-person scenes. Inference-time conditions are outlined in light blue.

et al., 2025a) use dual-stream architectures with cross-modal interaction. MOVA (Team et al., 2026) scales joint generation with a Mixture-of-Experts design. LTX-2 (HaCohen et al., 2026) uses an asymmetric dual-stream transformer, and daVinci-MagiHuman (Chern et al., 2026) uses a single-stream transformer without explicit cross-attention. These models still control the generated audio primarily through text, so emotion, rhythm, and prosody are inferred from the prompt. We keep the text prompt from the backbone and additionally inject driving speech, so that linguistic timing comes from the utterance while vocal identity is determined by the on-screen speaker.

3 Method

We propose an audio-driven human video generation model that restores a voice consistent with the visible speaker while retaining the linguistic content of the driving utterance. Given a driving utterance $\mathbf{x}^d$, a reference image $\mathbf{I}$, and an optional text prompt c, the model jointly generates an audio track $\mathbf{y}^a$ and a video $\mathbf{y}^v$. The generated audio should retain the linguistic and time-varying driving cues of $\mathbf{x}^d$ without copying its source-speaker timbre, and should stay close to the original paired recording of the on-screen speaker in $\mathbf{I}$. The generated video should preserve the visual identity in $\mathbf{I}$ and remain synchronized with $\mathbf{y}^a$. For two-person conversations, each driving stream can be bound to a designated visible speaker. The overall architecture is illustrated in Fig. 2.

3.1 Preliminaries

We build on the pretrained Ovi joint audio-video backbone (Low et al., 2025) and inherit its dual-stream flow-matching objective (Lipman et al., 2023). Let $m \in \{a, v\}$ denote the audio or video modality. Given a paired sample and text prompt c, clean latents $\mathbf{z}_0^m$ are linearly interpolated with Gaussian noise $\boldsymbol{\epsilon}^m$ as $\mathbf{z}_t^m = (1-t)\mathbf{z}_0^m + t\boldsymbol{\epsilon}^m$, with target velocity $\mathbf{v}_t^{m,*} = \boldsymbol{\epsilon}^m - \mathbf{z}_0^m$. The dual-stream model $\mathbf{v}_\Theta(\mathbf{z}_t^a, \mathbf{z}_t^v, t, c)$ jointly predicts both velocities and is optimized by

$$\mathcal{L}_{\text{fm}} = \mathbb{E}\left[\left\|\hat{\mathbf{v}}_t^a - \mathbf{v}_t^{a,*}\right\|_2^2 + \lambda_v \left\|\hat{\mathbf{v}}_t^v - \mathbf{v}_t^{v,*}\right\|_2^2\right], \tag{1}$$

where λ_v balances the audio and video terms. Our contributions add driving-audio conditioning and masked cross-modal interaction on top of this inherited process.

3.2 Audio Injection with Timbre-Perturbed Conditioning

To incorporate driving-audio information, we introduce an injection module based on Whisper (Radford et al., 2023) features. Given a driving clip, a pretrained Whisper encoder extracts high-level

tokens $\mathbf{R}$, which condition the audio generation stream. Whisper is more content-oriented than an audio VAE latent, but it may still retain speaker information; timbre-perturbed training (Sec. 3.4) supplies the pressure against copying the driving-speaker timbre.

The extracted tokens are injected into the audio generation stream through cross-attention:

$$\widetilde{\mathbf{H}}_t^a = \mathbf{H}_t^a + \mathrm{CrossAttn}(\mathbf{H}_t^a, \mathbf{R}), \tag{2}$$

where $\mathbf{H}_t^a$ denotes the hidden tokens of the noisy audio latent at timestep t. We use global cross-attention rather than window-based attention, so each generated audio token can attend to all driving-audio tokens and capture utterance-level content and temporal structure.

Following Ovi (Low et al., 2025), we align the temporal resolutions of the driving tokens and the noisy audio latent with scaled RoPE. When the driving audio contains N_r tokens and the noisy audio latent contains N_a tokens, we scale the RoPE frequencies of the driving stream by $\frac{N_a}{N_r}$, or equivalently evaluate the driving RoPE at

$$\tilde{p}_i^r = \frac{N_a}{N_r} p_i^r, \tag{3}$$

where $\tilde{p}_i^r$ is the effective RoPE position of the i-th driving token. In two-speaker cases, each driving stream is encoded independently and packed as a separate Whisper sequence; both streams condition the shared audio denoising branch, while speaker-to-face assignment is imposed by the audio–video masks below rather than by the Whisper injection.

3.3 Spatio-Temporal Masking for Two-Person Scenarios

To support two-person conversational scenes, we introduce a spatio-temporal masking mechanism for voice-face binding. At inference, the association between each driving stream and visible speaker can be specified by the user or obtained automatically. We detect identities in the conditioning first frame using InsightFace (Guo et al., 2018; Deng et al., 2019), partition the entire frame according to the detected identities, and keep the resulting regions fixed throughout generation. Partitioning the full frame, rather than only the mouth or face, gives each person a broader editable region for facial motion and body-related dynamics.

The temporal mask marks speaking intervals in each driving stream, obtained with Silero VAD (Silero Team, 2024). Given K speakers, let $\mathcal{S}_k$ denote the video tokens assigned to the k-th person, and let $\mathcal{A}_k$ denote the generated-audio tokens in that person's speaking intervals. We define a spatio-temporal attention mask $\mathbf{M}$ between video and audio tokens as

$$\mathbf{M}_{ij} = \begin{cases} 1, & \text{if } i \in \mathcal{S}_k \text{ and } j \in \mathcal{A}_k, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where i indexes video tokens and j indexes tokens of the noisy audio latent. Audio tokens outside $\bigcup_k \mathcal{A}_k$ are treated as inactive (silence) and are not bound to a speaker-specific region.

Before the attention softmax, the binary mask is converted into an additive bias:

$$\mathbf{B}_{ij} = \begin{cases} 0, & \text{if } \mathbf{M}_{ij} = 1, \\ -\infty, & \text{if } \mathbf{M}_{ij} = 0. \end{cases} \tag{5}$$

Disallowed audio–video pairs therefore receive zero attention probability. The two-person training clips exclude substantial overlapping speech, as detailed in Appendix A.

The mask is applied to the bidirectional audio-video cross-attention over the noisy latents, not to the Whisper injection:

$$\widetilde{\mathbf{H}}_t^v = \mathbf{H}_t^v + \mathrm{MaskedCrossAttn}\left(\mathbf{H}_t^v, \mathbf{H}_t^a; \mathbf{M}\right), \tag{6}$$

$$\widetilde{\mathbf{H}}_t^a = \mathbf{H}_t^a + \mathrm{MaskedCrossAttn}\left(\mathbf{H}_t^a, \mathbf{H}_t^v; \mathbf{M}^\top\right), \tag{7}$$

where $\mathbf{H}_t^v$ and $\mathbf{H}_t^a$ are the hidden tokens of the video and audio branches at timestep t. $\mathbf{M}$ is used for video-to-audio attention and $\mathbf{M}^\top$ for audio-to-video attention. Each person's visual tokens therefore interact only with the speech tokens assigned to that person.

3.4 Training Strategy

We use SeedVC (Liu, 2024) to replace the speaker timbre of each training audio while retaining its speech content, creating a controlled mismatch between the driving audio and the original speaker in the paired video. The input-target timbre gap discourages copying of the driving speaker and supervises the model to generate a voice close to that original speaker, using the visual context as the identity cue. Whisper supplies a higher-level driving representation; SeedVC supplies the training pressure against source-timbre copying.

We adopt a two-stage training strategy. The first stage uses single-person audio-video data so that the injection module can learn from timbre-mismatched conditions without multi-speaker interactions. The second stage uses two-person data to learn voice-face binding under the proposed masks. During both stages we optimize the flow-matching objective. To preserve the pretrained audio-video prior, we fully fine-tune the newly introduced Whisper cross-attention layers and the pretrained audio-video cross-modal attention layers (now with masking), and apply LoRA (Hu et al., 2022) to the remaining pretrained modules.

4 Experiments

4.1 Experimental Settings

Dataset We collect 250K single-person talking video clips, mainly from OpenHumanVid (Li et al., 2025), and 60K dual-person conversation clips from publicly available videos, covering diverse speaker identities and interaction scenarios. To prevent the model from copying the driving timbre, we apply SeedVC (Liu, 2024) with target timbres sampled from an age- and gender-balanced library of 40 voices. This produces mismatched audiovisual pairs in which speech content is preserved while the driving timbre no longer corresponds to the on-screen speaker. We use the single-person data to learn from timbre-mismatched conditions and the dual-person data to learn voice-face binding. For evaluation, we adopt EMTD (Meng et al., 2024) for the single-person setting and InterActEyes (Zhong et al., 2025) for the dual-person setting. The training and evaluation sets are speaker-identity-disjoint: no person identity appears in both splits. Detailed data construction and filtering are provided in Appendix A.

Evaluation Metrics We evaluate our method from three perspectives: voice conversion quality, video generation quality, and lip synchronization. For voice conversion, we compute Speaker Embedding Cosine Similarity (SECS) using ECAPA-TDNN (Desplanques et al., 2020) between the generated audio and the original paired recording of the on-screen speaker, which is used as the ground-truth vocal identity. We further employ a pretrained wav2vec2-large model (Baevski et al., 2020) fine-tuned for age and gender prediction (Burkhardt et al., 2023) to derive Gender Accuracy (G-ACC) and Age Mean Squared Error (A-MSE), measuring whether the generated voice preserves the demographic attributes of that original speaker. We report Word Error Rate (WER) and Character Error Rate (CER) from a HuBERT-based (Hsu et al., 2021) automatic speech recognition model to evaluate speech intelligibility. For video quality, we adopt Fréchet Inception Distance (FID) (Heusel et al., 2017), Fréchet Video Distance (FVD) (Unterthiner et al., 2018), and FVD on cropped face regions (F-FVD). For lip synchronization, we employ the SyncNet (Chung & Zisserman, 2016) confidence score (Sync-C) and distance (Sync-D). The perceptual study in Sec. 4.3 evaluates timbre–identity consistency and overall audiovisual coherence.

Implementation Details Our model inherits from Ovi (Low et al., 2025), a pretrained joint video-audio generation model. During training, we fully optimize the cross-modal attention layers and the driving-audio cross-attention modules, while adapting the remaining network parameters with LoRA (Hu et al., 2022). We train the model using the AdamW optimizer with a learning rate of 1×10^{-5} and a batch size of 64.

4.2 Quantitative Evaluation

We evaluate our method against state-of-the-art approaches on voice conversion quality and talking-video generation. For all voice conversion evaluations, the original audio that matches the on-screen

Table 1: Comparison of voice conversion methods on EMTD.

Method	SECS ↑	G-ACC ↑	A-MSE ↓	WER ↓	CER ↓
FVMVC	0.1513	0.95	193.63	0.7372	0.4089
DreamVoice	<u>0.1721</u>	<u>1.00</u>	<u>183.28</u>	<u>0.1638</u>	<u>0.0653</u>
Ours	**0.2946**	**1.00**	**151.71**	**0.0943**	**0.0348**

Table 3: Comparison on the single-person EMTD dataset.

Method	FID ↓	FVD ↓	F-FVD ↓	Sync-C ↑	Sync-D ↓
FantasyTalking	39.16	369.04	364.75	4.16	10.95
OmniAvatar	33.63	365.75	314.31	5.91	9.07
AnyTalker	29.67	349.35	320.87	**8.25**	**7.23**
MultiTalk	32.27	374.66	299.62	7.61	7.67
DreamVoice + MultiTalk	32.50	378.56	**289.73**	7.45	7.64
Ours	**29.22**	**348.38**	332.62	7.67	8.06

speaker is used as the ground truth. We perturb its timbre with SeedVC (Liu, 2024) to construct identity-mismatched driving audio, and then evaluate whether each method restores a voice that is consistent with the visual identity and close to the original speaker distribution, while preserving the original speech content.

Comparison with Voice Conversion Methods Since cross-modal voice conversion conditioned on visual identity remains relatively underexplored, we compare our method with two representative baselines: FVMVC (Sheng et al., 2023), a face-guided voice conversion method, and DreamVoice (Hai et al., 2024), a text-guided voice conversion method. To adapt DreamVoice to our visually conditioned setting, we use a multimodal large language model to generate a descriptive text prompt characterizing the target speaker's expected vocal attributes from the input portrait, and then use this description as the prompt for DreamVoice. All methods are evaluated on EMTD using the same input speech content and target identity conditions whenever applicable.

Table 1 compares our method with existing voice conversion baselines on EMTD. Our method obtains the highest SECS and the lowest A-MSE, WER, and CER, and ties DreamVoice on G-ACC at 1.00. The higher SECS and lower A-MSE indicate closer agreement with the original paired voice and its demographic attributes, while the lower WER and CER show stronger preservation of speech intelligibility.

Comparison with Talking Video Methods For talking-video evaluation, all methods use the same identity-mismatched driving audio. We compare our method with FantasyTalking (Wang et al., 2025b), OmniAvatar (Gan et al., 2025), AnyTalker (Zhong et al., 2025), and MultiTalk (Kong et al., 2025), as well as a cascaded *DreamVoice + MultiTalk* baseline that first converts the mismatched audio to the target timbre using DreamVoice and then uses the converted audio to drive MultiTalk. Most baselines use a Wan2.1 I2V-14B (Wan et al., 2025) backbone, whereas our video branch uses the more compact Wan2.2 TI2V-5B (Wan et al., 2025).

Table 2: Comparison on the dual-person InterActEyes dataset.

Method	FID ↓	FVD ↓	Sync-C ↑
MultiTalk	60.32	424.51	**5.46**
AnyTalker	**45.38**	415.59	<u>5.09</u>
Ours	<u>46.07</u>	**398.75**	4.79

On the single-person EMTD benchmark, our method obtains the best FID and FVD and competitive face-region quality and synchronization. On the dual-person InterActEyes benchmark, it achieves the best FVD and the second-best FID, while its Sync-C is lower than those of MultiTalk and AnyTalker. The results therefore show a competitive visual-quality and synchronization trade-off while adding timbre harmonization. Notably, the cascaded *DreamVoice + MultiTalk* pipeline yields degraded synchronization compared with directly using mismatched audio, further motivating the need to address timbre harmonization and talking-video generation within a unified framework.

4.3 Perceptual User Study

We conduct a perceptual study on 12 randomly selected EMTD test videos with 21 participants. For each sample, results from our method and the DreamVoice+MultiTalk cascade are displayed side by side with randomized left-right order and can be replayed. Participants independently assess (1) timbre–identity consistency and (2) overall audiovisual coherence, choosing our method, the cascade, or *Comparable* for each criterion. This yields 252 responses per criterion. As shown in Table 4, our method is selected in 64.29% of responses for timbre–identity consistency and 53.97% for overall audiovisual coherence, compared with 26.98% and 30.56% for the cascade. The complete protocol is provided in Appendix B.

Table 4: Preference distribution in the perceptual user study on EMTD.

Preference	Timbre–Identity ↑	A/V Coherence ↑
Comparable	8.73%	15.48%
DreamVoice + MultiTalk	26.98%	30.56%
Ours	**64.29%**	**53.97%**

4.4 Qualitative Evaluation

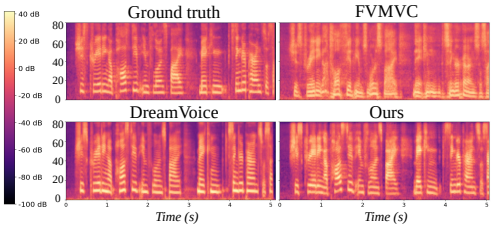

Figure 3: Mel-spectrogram comparison for qualitative evaluation.

Comparison with Voice Conversion Methods Figure 3 presents a qualitative comparison of mel-spectrograms generated by different voice conversion methods for the same linguistic content. The temporal and harmonic structures complement the WER, CER, and SECS results in Table 1. FVMVC exhibits visibly blurred spectral structures, whereas DreamVoice and our method retain clearer time-frequency patterns. A closer examination of speech with strong emotional expressions further reveals the advantage of our method. Specifically, our method preserves the linguistic content together with the emotion-related rhythm and prosodic dynamics of the source speech, while generating a target timbre that is harmonious with the newly rendered character. In comparison, the other two methods show greater degradation in preserving these characteristics. This suggests that our method more effectively disentangles key speech factors, including timbre, emotion, linguistic content, and prosody, enabling the target timbre to be adapted while retaining the desired content and expressive characteristics.

Comparison with Talking Video Methods Figure 4 visualizes key frames with enlarged mouth regions. The examples show temporally varying mouth shapes for the designated speakers and complement the SyncNet scores in Tables 3 and 2. Individual frames are not a substitute for sequence-level synchronization metrics. Although our Sync-C on InterActEyes is slightly lower than those of MultiTalk and AnyTalker, our method achieves comparable synchronization performance with a substantially more compact backbone for the video branch.

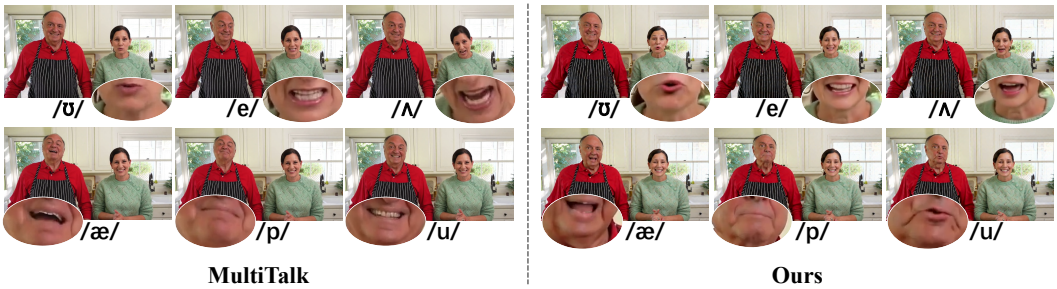

Figure 4: Talking video comparison for qualitative evaluation.

Table 5: Ablation study on driving-audio injection strategies.

Method	SECS ↑	G-ACC ↑	A-MSE ↓	WER ↓	CER ↓
VAE-Add	0.2630	0.98	190.18	0.3275	0.1503
VAE-ControlNet	0.2540	1.00	210.87	0.3343	0.1602
VAE-CrossAttn	0.2693	1.00	187.68	0.6040	0.3018
Ours	**0.2946**	**1.00**	**151.71**	**0.0943**	**0.0348**

4.5 Ablation Study

Driving-Audio Injection Strategy We ablate the driving-audio injection strategy. Our full model encodes the driving audio with a Whisper (Radford et al., 2023) encoder and injects the resulting tokens into the audio denoising network through cross-attention. We compare it with three alternatives: *VAE-Add*, which adds the audio VAE latent to the noisy audio latent; *VAE-ControlNet*, which injects VAE-encoded audio features through a ControlNet-like residual branch (Zhang et al., 2023a; Chen et al., 2024); and *VAE-CrossAttn*, which uses VAE-encoded audio features as cross-attention tokens. Table 5 shows that Whisper injection obtains the best or tied-best scores on all five metrics. The substantially lower WER and CER indicate that high-level Whisper features preserve linguistic content more effectively under timbre perturbation than VAE latents, which retain low-level acoustic details that can entangle timbre, noise, and prosody.

Spatio-Temporal Masking Mechanism We further ablate the proposed spatio-temporal masking mechanism for two-person voice-face binding (Table 6). In the *Global-Audio* variant, all driving streams are mixed and injected as a single global condition without speaker-region association, so all visible faces are driven by the same audio. We do not report Sync-C for this variant, because driving every speaker with the same audio inflates the score and does not measure speaker-audio correspondence. In the *ID-Audio Concatenation* variant, we concatenate each speaker's face token with the corresponding driving-audio tokens. This provides an implicit identity cue, but all tokens remain globally visible, which is insufficient to prevent cross-speaker leakage. Our full model restricts each generated-audio interval to its corresponding spatial region during the designated speaking interval. Global-Audio obtains a slightly better FID, while the full model improves FVD and Sync-C.

Table 6: Ablation study on the spatio-temporal masking mechanism.

Method	FID ↓	FVD ↓	Sync-C ↑
Global-Audio	**44.93**	572.27	–
ID-Audio Concat.	46.56	439.47	3.57
Ours	46.07	**398.75**	**4.79**

4.6 Limitations

Our experiments focus on short, stable single-person and two-person scenes with mostly non-overlapping speech. Because the spatial partition is derived from the conditioning frame and remains fixed, large movements or subjects exchanging positions can disrupt the correspondence between audio streams and visual regions. Overlapping speech, off-screen speakers, and errors from face analysis or voice activity detection are also outside the evaluated setting. The current clips are approximately five seconds long, so long-horizon timbre consistency remains unverified. The mismatch protocol is constructed with SeedVC at both training and test time.

5 Conclusion

We study audiovisual incongruity in audio-driven human video generation, where lip synchronization alone does not make the generated voice consistent with the visible speaker. Our framework combines a Whisper-based driving-audio injection module with timbre-perturbed training, so that linguistic content is retained from the driving utterance while the generated timbre is pulled toward the original paired recording of the on-screen speaker. A spatio-temporal masking mechanism supports explicit voice-face association in two-person conversations. Experiments on speaker-identity-disjoint single-person and dual-person benchmarks show closer agreement with the paired original voice, together with competitive visual quality and synchronization.

AI USE STATEMENT

In this work, we used generative AI tools to generate text descriptions of the training clips and prompts for the DreamVoice baseline. Gemini 2.5 Pro and MiMo-V2.5 generated the semantic descriptions used as text prompts inherited from the Ovi backbone. For two-person clips, the prompts were written so as not to reveal the speaking order.

Because DreamVoice is text-guided, we used MiMo-V2.5 to describe each speaker's vocal attributes based on the corresponding conditional image. Each description served as the prompt for DreamVoice to convert the corresponding utterance. The converted utterances were used in the EMTD voice-conversion comparison and, after being used to drive MultiTalk, in the EMTD talking-video comparison, the mel-spectrogram examples, and the perceptual study.

Generative AI tools were not used to develop the conceptual framework, formulate mathematical claims, write proofs, propose or refine hypotheses, design the experiments, implement the method, select hyperparameters or loss functions, design the masking mechanism, translate text, analyze user-study responses, interpret quantitative results, or produce any reported numbers. We also used generative AI tools to edit the manuscript's language and organization. The authors drew the figures.

We reviewed the generated training descriptions, DreamVoice prompts, and edited manuscript. We take responsibility for the final content of this work, including the text, claims, and data produced with the assistance of generative AI.

ETHICS STATEMENT

The ability to modify voices and synthesize identity-consistent human videos can be misused for impersonation, deceptive media, or unauthorized dubbing. Responsible deployment should require appropriate authorization from depicted and recorded individuals and should incorporate provenance or watermarking where possible. Training on appearance–voice correlations may also amplify demographic stereotypes present in the data. The perceptual study is reported only through aggregate preference judgments; data processing and the study protocol are described in the appendix.

REPRODUCIBILITY STATEMENT

The model formulation and conditioning mechanisms are specified in Sec. 3. Training data construction, speaker-identity-disjoint evaluation, timbre perturbation, and the perceptual-study protocol are documented in Appendix A and Appendix B.

REFERENCES

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, 2020.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

Felix Burkhardt, Johannes Wagner, Hagen Wierstorf, Florian Eyben, and Björn Schuller. Speech-based age and gender prediction with transformers. In *15th ITG Conference on Speech Communication*, 2023.

Zexin Cai, Henry Li Xinyuan, Ashi Garg, Leibny Paola García-Perera, Kevin Duh, Sanjeev Khudanpur, Matthew Wiesner, and Nicholas Andrews. GenVC: Self-supervised zero-shot voice conversion. *arXiv preprint arXiv:2502.04519*, 2025.

Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. PIXART-δ: Fast and Controllable Image Generation with Latent Consistency Models. *arXiv preprint arXiv:2401.05252*, 2024.

Ethan Chern, Hansi Teng, Hanwen Sun, Hao Wang, Hong Pan, Hongyu Jia, Jiadi Su, Jin Li, Junjie Yu, Lijie Liu, et al. Speed by simplicity: A single-stream architecture for fast audio-video generative foundation model. *arXiv preprint arXiv:2603.21986*, 2026.

J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254*, 2019.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*, 2020.

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Qijun Gan, Ruizi Yang, Jianke Zhu, Shaofei Xue, and Steven Hoi. OmniAvatar: Efficient audio-driven avatar video generation with adaptive body animation. *arXiv preprint arXiv:2506.18866*, 2025.

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, et al. Wan-s2v: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025.

Google DeepMind. Veo: a text-to-video generation system. Technical report, Google DeepMind, 2025.

Jia Guo, Jiankang Deng, et al. InsightFace: 2D and 3D face analysis project, 2018.

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.

Jiarui Hai, Karan Thakkar, Helin Wang, Zengyi Qin, and Mounya Elhilali. DreamVoice: Text-guided voice conversion. *arXiv preprint arXiv:2406.16314*, 2024.

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. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Teng Hu, Zhentao Yu, Guozhen Zhang, Zihan Su, Zhengguang Zhou, Youliang Zhang, Yuan Zhou, Qinglin Lu, and Ran Yi. Harmony: Harmonizing audio and video generation through cross-task synergy. *arXiv preprint arXiv:2511.21579*, 2025.

Yubo Huang, Weiqiang Wang, Sirui Zhao, Tong Xu, Lin Liu, and Enhong Chen. Bind-Your-Avatar: Multi-talking-character video generation with dynamic 3D-mask-based embedding router. *arXiv preprint arXiv:2506.19833*, 2025.

Youngjoon Jang, Ji-Hoon Kim, Junseok Ahn, Doyeop Kwak, Hong-Sun Yang, Yoon-Cheol Ju, Il-Hwan Kim, Byeong-Yeol Kim, and Joon Son Chung. Faces that speak: Jointly synthesising talking face and speech from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *Proc. 2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018.

Takuhiro Kaneko and Hirokazu Kameoka. CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. Let them talk: Audio-driven multi-person conversational video generation. *arXiv preprint arXiv:2505.22647*, 2025.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. OpenHumanVid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jingyi Li, Weiping Tu, and Li Xiao. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, Chao Liang, Yuan Zhang, and Jingtuo Liu. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Songting Liu. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*, 2024.

Songxiang Liu, Yuewen Cao, Dan Su, and Helen Meng. DiffSVC: A diffusion probabilistic model for singing voice conversion. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.

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

Hsiao-Han Lu, Shao-En Weng, Ya-Fan Yen, Hong-Han Shuai, and Wen-Huang Cheng. Face-based voice conversion: Learning the voice behind a face. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. EchoMimicV2: Towards striking, simplified, and semi-body human animation. *arXiv preprint arXiv:2411.10061*, 2024.

Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T. Freeman, Michael Rubinstein, and Wojciech Matusik. Speech2Face: Learning the face behind a voice. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

OpenAI. Sora 2 system card. Technical report, OpenAI, 2025.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Sergeevich Kudinov, and Jiansheng Wei. Diffusion-based voice conversion with fast maximum likelihood sampling scheme. In *International Conference on Learning Representations*, 2022.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. AutoVC: Zero-shot voice style transfer with only autoencoder loss. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2019.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 2023.

Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Yan Rong and Li Liu. Seeing your speech style: A novel zero-shot identity-disentanglement face-based voice conversion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23), 2025.

Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Zheng-Yan Sheng, Yang Ai, Yan-Nian Chen, and Zhen-Hua Ling. Face-driven zero-shot voice conversion with memory-based face-voice alignment. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

Silero Team. Silero VAD: Pre-trained enterprise-grade voice activity detector, 2024.

Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, 2023.

Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-jin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (ToG)*, 43(4), 2024.

OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, et al. Mova: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026.

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*, 2024.

Shuyuan Tu, Yueming Pan, Yinming Huang, Xintong Han, Zhen Xing, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Stableavatar: Infinite-length audio-driven avatar video generation. *arXiv preprint arXiv:2508.08248*, 2025.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

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.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. UniVerse-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025a.

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. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025b.

Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021.

Zhenzhi Wang, Jiaqi Yang, Jianwen Jiang, Chao Liang, Gaojie Lin, Zerong Zheng, Ceyuan Yang, Yuan Zhang, Mingyuan Gao, and Dahua Lin. Interacthuman: Multi-concept human animation with layout-aligned audio conditions. *arXiv preprint arXiv:2506.09984*, 2025c.

Shao-En Weng, Hong-Han Shuai, and Wen-Huang Cheng. Zero-shot face-based voice conversion: Bottleneck-free speech disentanglement in the real-world scenario. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11), 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025.

Jixun Yao, Yuguang Yang, Yi Lei, Ziqian Ning, Yanni Hu, Yu Pan, Jingjing Yin, Hongbin Zhou, Heng Lu, and Lei Xie. Promptvc: Flexible stylistic voice conversion in latent space driven by natural language prompts. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

Guozhen Zhang, Zixiang Zhou, Teng Hu, Ziqiao Peng, Youliang Zhang, Yi Chen, Yuan Zhou, Qinglin Lu, and Limin Wang. Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions. *arXiv preprint arXiv:2511.03334*, 2025a.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023a.

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. S^3FD: Single shot scale-invariant face detector. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Tian-Hao Zhang, Jiawei Zhang, Jun Wang, Xinyuan Qian, and Xu-Cheng Yin. FaceSpeak: Expressive and high-quality speech synthesis from human portraits of different styles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025b.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b.

Zhizhou Zhong, Yicheng Ji, Zhe Kong, Yiying Liu, Jiarui Wang, Jiasun Feng, Lupeng Liu, Xiangyi Wang, Yanjia Li, Yuqing She, et al. Anytalker: Scaling multi-person talking video generation with interactivity refinement. *arXiv preprint arXiv:2511.23475*, 2025.

Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei 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*, 2021.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6), 2020.

A Dataset Collection and Processing

A.1 Single-Person Talking Videos

Our single-person training data is mainly collected from OpenHumanVid (Li et al., 2025) and re-processed through a multi-stage pipeline to obtain high-quality talking-video clips. We first perform scene segmentation and preliminary low-quality filtering. For each remaining clip, we apply the S3FD face detector (Zhang et al., 2017) and retain only clips in which exactly one face is consistently visible. We then combine voice activity detection (VAD) with active speaker detection (ASD) to identify speaking intervals and verify that the active speaker is present on screen throughout the corresponding speech segments. To further ensure reliable audio-visual correspondence, we use SyncNet (Chung & Zisserman, 2016) to compute the lip-sync confidence score and discard clips with scores below 1.5. Finally, we employ a multimodal large language model to generate semantic descriptions for the retained clips; these descriptions are used as the text prompt c inherited from the Ovi backbone. This process yields approximately 250K clips. Each clip contains 121 frames at 24 fps, corresponding to about 5.04 seconds, and is processed at 720p resolution.

A.2 Two-Person Conversational Videos

The two-person data is collected from publicly available online videos and processed through a dedicated conversational-data pipeline. We first apply face detection and retain only segments that consistently contain exactly two visible faces. Demucs (Défossez et al., 2019) is then used to suppress background music and other non-speech components. Next, pyannote speaker-diarization-3.1 (Bredin et al., 2020) identifies the speaking intervals of the two speakers and the timestamps at which the active speaker changes. We apply a sliding temporal window around these transition points, ensuring that each extracted clip contains speech from both speakers and at least one conversational turn transition. To establish the correspondence between voices and visible identities, we remove intervals with substantial overlapping speech and use the remaining diarization results to construct a speaker-specific audio stream for each of the two speakers. This yields two audio streams and two face tracks for every clip. We compute SyncNet (Chung & Zisserman, 2016) lip-sync scores for all four audio–face pairs and determine the assignment under a one-to-one constraint: the valid correspondence must be either $(a_1, f_1), (a_2, f_2)$ or $(a_1, f_2), (a_2, f_1)$, rather than assigning both audio streams to the same face. The assignment with the higher aggregate lip-sync score is retained, while samples with unreliable correspondence are discarded. Finally, a multimodal large language model annotates the retained clips using a prompt that does not reveal the speaking order of the two identities, thereby preventing order information from leaking into the text condition. The resulting training set contains approximately 60K two-person conversational clips and is used to learn voice–face binding. The EMTD (Meng et al., 2024) and InterActEyes (Zhong et al., 2025) evaluation sets are speaker-identity-disjoint from the training data: no person identity appears in both training and evaluation.

A.3 Timbre Library and Perturbation

We manually construct a library of 40 perceptually distinct timbres. The library covers four age groups (children, teens, adults, and elderly) and two gender groups (male and female), with five distinct timbres in each age-gender group. Each timbre is represented by a 5-second reference audio clip. During training, we randomly sample a target timbre from this library and use SeedVC (Liu, 2024) to perturb the original speech. This preserves the linguistic content while creating an input-target timbre mismatch, preventing the model from learning the shortcut of copying the driving speaker's timbre. The original unperturbed recording remains the training target and, at test time, the ground-truth vocal identity.

B User Study Protocol

We conduct a perceptual user study with 21 participants and 12 randomly selected videos from the EMTD test set. Each trial displays the outputs of our method and the DreamVoice (Hai et al., 2024)+MultiTalk (Kong et al., 2025) cascade side by side. The left-right placement of the two methods is randomized, and participants may replay the videos before voting.

Participants evaluate two criteria independently: (1) timbre–identity consistency, which assesses whether the generated voice matches the visible speaker, and (2) overall audiovisual coherence, which assesses the harmony of the generated audio and video as a whole. For each criterion, participants choose *Ours*, *DreamVoice+MultiTalk*, or *Comparable*. Since every participant evaluates all 12 videos, each criterion contains $21 \times 12 = 252$ responses. The preference distribution is reported in Table 4 of the main paper.