M^{2}TALK: AUDIO-DRIVEN MULTI-PERSON MULTI-SHOT CONVERSATIONAL VIDEO GENERATION VIA SPEAKER-IDENTITY CONDITIONING

Anonymous authors
Paper under double-blind review

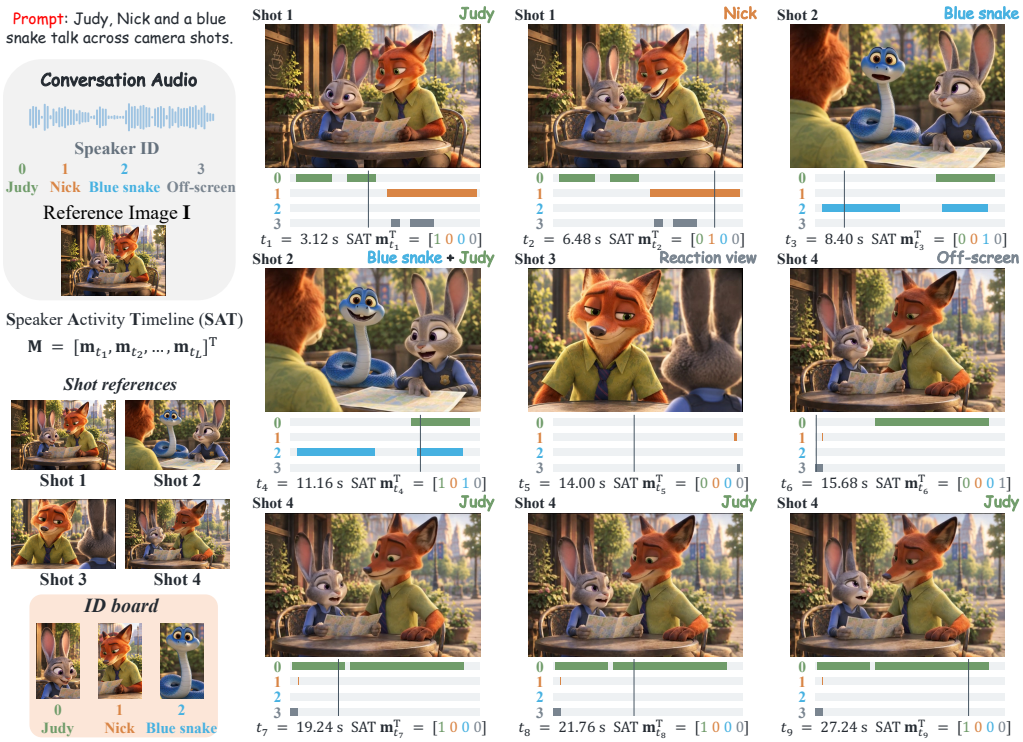


Figure 1: We propose M^{2}Talk, a framework for audio-driven multi-person conversational video generation. Given conversation audio, a prompt, a reference image, and a supplied Speaker Activity Timeline (SAT), M^{2}Talk generates interactions with audio-aligned lip motions and consistent speaker assignment. Adding *shot references* and an *ID board* enables multi-shot generation, including participants who appear only in later shots.

ABSTRACT

Audio-driven human video generation has progressed from animating single person to synthesizing multi-person conversations. Prevailing multi-person talking video frameworks rely heavily on spatially grounded audio binding, where speech features are hard-routed to target visual regions via spatial masks. However, this rigid spatial coupling fundamentally limits natural human interactions, frequently causing binding failures under dynamic movements, suppressing non-verbal listener dynamics, and failing during off-screen speech or shot transitions. To overcome these boundaries, we propose M^{2}Talk, a framework for multi-person multi-shot conversational video generation built on speaker-identity conditioning. Our key insight is to decouple *who is speaking and when* from visual localization. Specifically, Speaker-Aware Audio Modulation (SAM) encodes a supplied Speaker Activity Timeline (SAT) into time-aligned Speaker-Aware Embeddings (SAEs) that modulate acoustic features without per-person spatial masks. Temporal Context Propagation (TCP) preserves visual coherence across consecutive

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.

generation chunks, while Multi-Reference Visual Conditioning (MVC) supplies shot reference images and ID board via dedicated condition slots to support multi-shot transitions. Trained on diverse natural conversational videos, M^{2}Talk effectively handles complex turn-taking, off-screen speech, and multi-shot, achieving state-of-the-art performance on audio-driven video generation benchmarks.
Project page: https://anonymous.4open.science/w/M2Talk-F011.

1 INTRODUCTION

Audio-driven human video generation aims to synthesize human-centric videos in which facial expressions and body movements are coordinated with input speech. Recent progress has expanded its scope from animating individual faces to producing expressive talking bodies (Meng et al., 2024; Lin et al., 2025; Wang et al., 2025a; Gan et al., 2025; Jiang et al., 2025; Gao et al., 2025; Yang et al., 2025) and multi-person interactions. These advances bring the field closer to generating complete conversational scenes, where speech, each person's behavior, and visual presentation must remain coherent as a conversation unfolds.

Extending single-person animation to multi-person conversations introduces an audio–person binding problem: each audio stream must be associated with its intended person (*i.e.*, speaker) without inducing speech-related lip movements in non-speaking persons. To specify *who is speaking and when*, existing methods commonly utilize spatial information to associate each audio stream with its target person's temporally aligned visual regions. HunyuanVideo-Avatar (Chen et al., 2025b) and InterActHuman (Wang et al., 2025c) associate audio conditions with person layouts or masks. Bind-Your-Avatar (Huang et al., 2025) and MultiTalk (Kong et al., 2025) dynamically localize people to route audio, while AnyTalker (Zhong et al., 2025) pairs speech with facial features and aggregates the resulting attention outputs using face masks. Despite differences in their implementations, these approaches share a spatially grounded binding strategy: using each person's location to determine where the associated speech should influence the generated video.

However, existing methods based on spatially grounded audio binding face several limitations in natural conversational scenes. First, maintaining audio–person correspondence depends on accurately localizing the intended speaker as the scene evolves. For example, a person may continue speaking while changing position, turning their head, or becoming partially occluded by another participant. If the region used for audio conditioning does not follow these changes, speech guidance may miss the intended speaker or affect another person. Second, even when localized audio control produces synchronized facial motion, the conversation as a whole may still appear unnatural because body movements and listener reactions are insufficiently coordinated. For instance, speakers may accompany their utterances with head movements and hand gestures, while silent listeners may nod, turn toward the speaker, or change expression without appearing to speak. Local speech control therefore needs complementary motion and interaction modeling to coordinate behavior beyond the directly conditioned regions. Third, multi-shot conversations may introduce speakers who are absent from the initial shot and its reference image. For example, in Fig. 1, the first shot shows Judy and Nick, while the blue snake appears in Shot 2 and joins the conversation. The binding mechanism must therefore accommodate newly appearing speakers and preserve correct audio–person correspondence despite changes in person visibility across shots.

Our key insight is to express *who is speaking and when* directly in the acoustic representation, thereby decoupling speech assignment from spatial visual masks. Instead of enforcing rigid spatial restrictions, we conditions the video generator on speaker-aware audio representations, allowing natural gestures and flexible listener dynamics to emerge organically. Based on this perspective, we introduce **M^{2}Talk**, a novel framework for audio-driven multi-person multi-shot conversational video generation built on **speaker-identity conditioning**.

To implement this idea, we introduce Speaker-Aware Audio Modulation (**SAM**). SAM encodes a supplied Speaker Activity Timeline (**SAT**), which specifies *who is speaking and when*, into a sequence of time-aligned Speaker-Aware Embeddings (**SAEs**). SAM uses these embeddings to modulate acoustic features before they enter the video generator, providing explicit speaker information without applying per-person spatial masks to audio-attention outputs. The resulting condition specifies speaking activity without prescribing each participant's facial or body motion, leaving speaking gestures and listener reactions to be learned from the conversational context. The SAT distinguishes

on-screen from off-screen speech, allowing the conditioning to account for participants entering or leaving the frame within a shot and ongoing speech across shots. To support temporal coherence across generation chunks, we introduce Temporal Context Propagation (**TCP**). TCP propagates trailing frames as context for the next chunk, keeping SAT aligned with the input audio.

A generation chunk can span multiple shots, each requiring its own reference image to specify the shot's composition. We therefore extend the backbone's single-reference conditioning to Multi-Reference Visual Conditioning (**MVC**), which incorporates *shot-specific reference images*. Moreover, participants who appear in later shots may be absent from the initial reference image. MVC adds an *ID board* of person crops across shots, supplying identity cues throughout generation.

Our main contributions are summarized as follows: (1) We formulate the task of audio-driven multi-person multi-shot conversational video generation and present M^{2}Talk to resolve its fundamental cross-modal challenges; (2) We rethink traditional spatial-masking paradigms and propose Speaker-Aware Audio Modulation (SAM) to achieve spatial-mask-free speaker conditioning; (3) Extensive experiments across talking-head, talking-body, and conversational benchmarks demonstrate our method's superiority in both speaker binding accuracy and synthesis quality.

2 Related Work

2.1 Audio-Driven Human Generation

Single-person generation. Audio-driven human generation has progressed from talking heads (Prajwal et al., 2020; Zhou et al., 2020; Zhang et al., 2023) to talking bodies with coordinated gestures (Meng et al., 2024; Lin et al., 2025; Wang et al., 2025a; Gan et al., 2025). FantasyTalking (Wang et al., 2025a) and InfiniteTalk (Yang et al., 2025) use audio cross-attention (Fig. 2(a)), while OmniAvatar (Gan et al., 2025) adopts additive conditioning (Fig. 2(b)). Further work explores semantic and cinematic control (Jiang et al., 2025; Gao et al., 2025) and long-duration synthesis (Gan et al., 2025; Yang et al., 2025).

Multi-person generation. Multi-person generation additionally requires correct audio–person correspondence while non-speaking participants listen. A common strategy grounds speech in person-specific visual regions (Chen et al., 2025b; Wang et al., 2025c). Representative approaches include MultiTalk (Kong et al., 2025) (Fig. 2(c)), Bind-Your-Avatar (Huang et al., 2025) (Fig. 2(d)), and AnyTalker (Zhong et al., 2025) (Fig. 2(e)); Playmate2 (Ma et al., 2025) further combines person-specific conditions. M^{2}Talk instead conditions audio on supplied speaker activity and speaker–person correspondence, without applying per-person spatial masks to audio-attention outputs (Fig. 2(f)); Section 3.2 details the formulation.

2.2 Multi-Shot Video Generation

Multi-shot video generation explores shot planning and reference-based control (Kara et al., 2025; Wang et al., 2025b; Zhang et al., 2025b), as well as visual memory for cross-shot consistency (Luo et al., 2025; Zhang et al., 2025a). MoCha and TalkCuts investigate speech-driven character and multi-shot generation (Wei et al., 2025; Chen et al., 2025a), while UnityShots and JoyAI-Echo-1.5 incorporate audiovisual conditioning and memory (Huang et al., 2026; Duan et al., 2026). TCR explicitly controls shot and dialogue timing (Liu et al., 2026), and MiniMax-H3 supports multimodal references and structured shot, speaker, and audio-reuse instructions (MiniMax, 2026). Our focus is correct speech assignment under changing speaker visibility: a silent listener should not appear to speak when an utterance continues off screen. M^{2}Talk addresses this setting by modulating acoustic features with supplied speaker activity, together with shot references and temporal context.

3 Method

3.1 Preliminaries

We adopt a DiT-based video diffusion model that takes a text prompt $\mathbf{T}$ and a reference image $\mathbf{I}$ as inputs. A 3D variational autoencoder (VAE) encodes and decodes video, while a text encoder and a CLIP image encoder (Radford et al., 2021) map $\mathbf{T}$ and $\mathbf{I}$ to $\mathbf{c}_{\text{text}}$ and $\mathbf{c}_{\text{image}}$, respectively. These

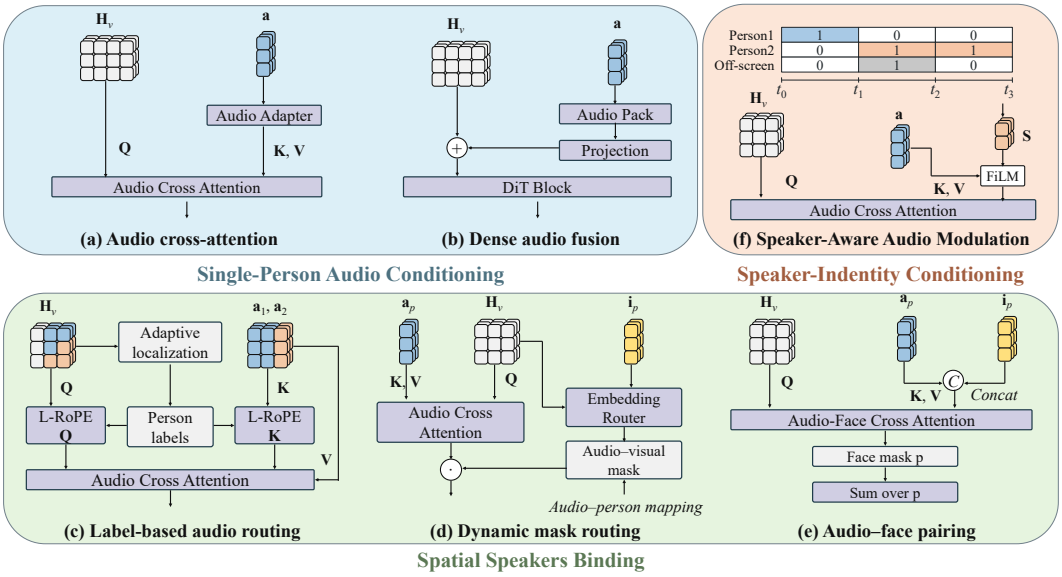


Figure 2: **Audio conditioning for single-person and conversational generation.** $\mathbf{H}_v$ denotes visual token features and $\mathbf{a}$ denotes acoustic features. For person p, $\mathbf{a}_p$ and $\mathbf{I}_p$ are paired audio and image features; p also indexes the face mask and aggregation branch. Bold $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ denote attention queries, keys, and values. In (f), one audio track contains both speakers, and the SAE sequence $\mathbf{S}$ modulates its features through FiLM.

two conditions are injected through decoupled cross-attention. For audio-driven generation, we additionally introduce a raw audio signal $\mathbf{a}_{\text{raw}}$. Wav2Vec (Baevski et al., 2020) extracts a sequence of audio features $\mathbf{a}_t^{\text{enc}}$ from $\mathbf{a}_{\text{raw}}$, where $t = 1, \ldots, L$ indexes positions aligned with the video-frame time grid. Each feature retains the encoder-layer axis and has d_a feature channels. Following Tian et al. (2024), we aggregate neighboring features within a temporal window of radius h:

$$\mathbf{a}_t = \left[\mathbf{a}_{t-h}^{\text{enc}}, \ldots, \mathbf{a}_{t+h}^{\text{enc}}\right]. \tag{1}$$

This operation introduces a local window axis of length $2h + 1$ while preserving the encoder-layer and feature-channel axes. We denote the resulting sequence of locally contextualized acoustic features by $\mathbf{a} = (\mathbf{a}_t)_{t=1}^{L}$.

Figure 3 provides an overview of M^{2}Talk. Given the text and image conditions above, Speaker-Aware Audio Modulation (SAM) incorporates speaker activity into the acoustic condition, Temporal Context Propagation (TCP) propagates recent visual context across video chunks, and Multi-Reference Visual Conditioning (MVC) provides identity and shot-level references. These components jointly condition the generation of each video chunk $\widehat{\mathbf{X}}^{(k)}$:

$$\widehat{\mathbf{X}}^{(k)} = \mathcal{G}_\theta\left(\mathbf{c}_{\text{text}}, \mathbf{c}_{\text{image}}; \mathbf{a}_{\text{cond}}^{(k)}, \mathbf{Z}_{\text{anc}}^{(k)}, \mathbf{I}_{\text{ref}}, \mathbf{I}_{\text{orig}}, \mathbf{m}_{\text{orig}}\right). \tag{2}$$

Here, $\mathcal{G}_\theta$ denotes the diffusion sampling and VAE decoding process parameterized by θ, with sampling randomness left implicit. The chunk-aligned audio condition $\mathbf{a}_{\text{cond}}^{(k)}$ is produced by SAM, while the clean anchor $\mathbf{Z}_{\text{anc}}^{(k)}$ unifies the initial motion frame and subsequent TCP context, as defined in Eq. 6. MVC provides the reference latents $\mathbf{I}_{\text{ref}}$, together with the backbone's original image condition $\mathbf{I}_{\text{orig}}$ and mask $\mathbf{m}_{\text{orig}}$.

3.2 Speaker-Aware Audio Modulation

For multi-person conversational generation, associating speech activity with the intended speakers is essential for coherent and controllable animation. Existing methods often rely on spatial localization or per-person masking to establish this correspondence. For example, MultiTalk (Kong et al., 2025) adaptively localizes individual people through reference-to-video self-attention and employs person

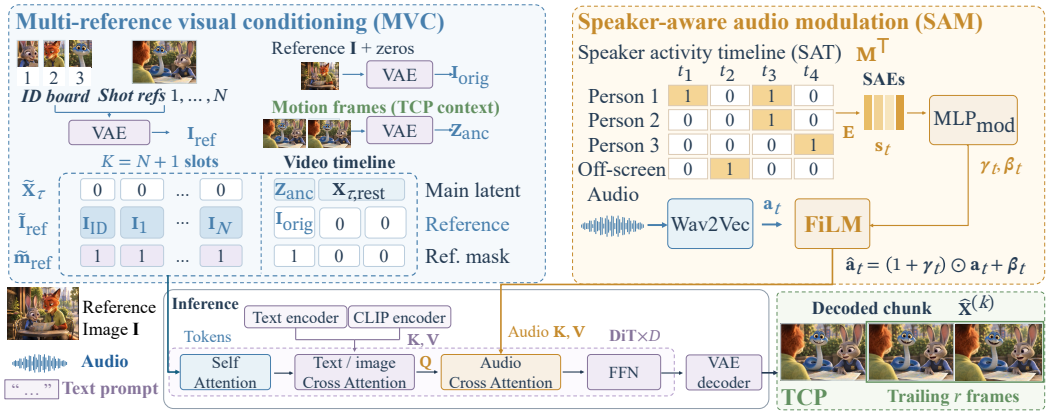


Figure 3: **Overview of M^{2}Talk.** SAM uses the supplied Speaker Activity Timeline (SAT) to modulate conversation audio. TCP propagates recent visual context to support chunk-wise continuation. MVC uses optional shot references and an ID board for multi-shot generation.

labels to rotate visual queries and audio keys using L-RoPE (Fig. 2(c)). Bind-Your-Avatar (Huang et al., 2025) predicts dynamic spatiotemporal masks during denoising to route audio conditions to their corresponding subjects (Fig. 2(d)). AnyTalker (Zhong et al., 2025) explicitly pairs encoded audio and facial features, applies shared cross-attention, and aggregates the resulting person-specific representations through face-mask weighting (Fig. 2(e)).

Despite their effectiveness, these spatially grounded binding strategies have inherent limitations. Because audio-person correspondence is enforced through spatial localization, the binding can become brittle when subjects undergo substantial head rotations or occlusions, causing localized speech control to fail. Moreover, the strong spatial constraints can restrict the natural dynamics of non-speaking subjects, often resulting in unnatural or overly static poses for listeners.

To address these limitations, we instead encode speaker identity and temporal activity directly into the audio condition, leveraging the given audio-person correspondence. While acoustic features capture the speech signal itself, they do not explicitly indicate which person should speak at a given time. SAM therefore converts a supplied Speaker Activity Timeline (SAT) into time-aligned Speaker-Aware Embeddings (SAEs), which explicitly encode speaker identity and activity and modulate the acoustic features before audio cross-attention (Fig. 2(f)). This design decouples audio-person binding from explicit spatial masks, enabling the audio condition to carry speaker-specific information throughout the generation process.

The supplied SAT is represented as $\mathbf{M} \in \{0,1\}^{L\times S}$, where the t-th row is $\mathbf{m}_t^\top$. The S activity channels consist of person slots ordered from left to right according to the persons in the reference image or *ID board*, together with one shared off-screen slot. The scalar $m_{t,s}$ indicates whether channel s is active at position t, with one denoting speech and zero denoting inactivity. Unused person slots are kept at zero. Multiple active channels represent overlapping speech, whereas $\mathbf{m}_t = \mathbf{0}$ denotes non-speech. The off-screen slot indicates unseen speech without requiring the identity of the unseen speaker to be specified.

SAM converts these discrete activity indicators into continuous speaker-aware conditions using a learnable embedding table $\mathbf{E} \in \mathbb{R}^{(S+1)\times d_s}$, where d_s denotes the SAE dimension. Each row $\mathbf{e}_s \in \mathbb{R}^{d_s}$ corresponds to an activity channel for $s = 1, \ldots, S$, while $\mathbf{e}_0$ represents the non-speech state. At each position t, the SAE $\mathbf{s}_t \in \mathbb{R}^{d_s}$ is obtained by summing the embeddings of all active channels:

$$\mathbf{s}_t = \begin{cases} \sum_{s=1}^{S} m_{t,s}\mathbf{e}_s, & \|\mathbf{m}_t\|_1 > 0, \\ \mathbf{e}_0, & \|\mathbf{m}_t\|_1 = 0. \end{cases} \tag{3}$$

The resulting SAE sequence $\mathbf{S} \in \mathbb{R}^{L\times d_s}$ consists of $\mathbf{s}_t^\top$ as its t-th row. Here, bold $\mathbf{S}$ denotes the SAE sequence, whereas scalar S denotes the number of activity channels. In our implementation, $S = 5$ and $d_s = d_a = 768$, yielding a 6×768 embedding table corresponding to four person slots, one off-screen slot, and the separate non-speech embedding.

We inject the speaker-aware condition into the acoustic features through Feature-wise Linear Modulation (FiLM) (Perez et al., 2018). Specifically, each SAE predicts scale and shift vectors $\boldsymbol{\gamma}_t, \boldsymbol{\beta}_t \in \mathbb{R}^{d_a}$ for the acoustic window centered at position t:

$$[\boldsymbol{\gamma}_t; \boldsymbol{\beta}_t] = \mathrm{MLP}_{\mathrm{mod}}\left(\mathrm{LN}(\mathbf{s}_t)\right), \qquad \widehat{\mathbf{a}}_t = (1 + \boldsymbol{\gamma}_t) \odot \mathbf{a}_t + \boldsymbol{\beta}_t. \tag{4}$$

Here, LN is layer normalization, $\mathrm{MLP}_{\mathrm{mod}}$ is a multilayer perceptron, and $\odot$ denotes element-wise multiplication. Scale and shift broadcast across the window and encoder-layer axes; zero initialization of the final modulation layer makes FiLM initially an identity transformation. The audio adapter uses temporal grouping and learned projections to map $\widehat{\mathbf{a}}_t$ onto the VAE's compressed time grid, producing the chunk-aligned condition $\mathbf{a}_{\mathrm{cond}}^{(k)}$ in Eq. 2 for audio cross-attention. An inactive visible-person channel indicates non-speech without constraining that person's body motion, facial expressions, or listener reactions.

3.3 Temporal Context Propagation

Extending conversational generation over long durations requires preserving appearance and motion across successive video chunks. Following chunk-wise continuation methods (Wan et al., 2025; Kong et al., 2025; Yang et al., 2025), TCP uses a short segment of generated video as the clean anchor for the next chunk. This segment carries recent spatiotemporal context to guide continued motion, complementing the image references and SAM's audio condition.

After decoding chunk $\widehat{\mathbf{X}}^{(k)}$, TCP re-encodes its trailing r frames as context for the next chunk:

$$\mathbf{C}_k = \mathrm{Tail}_r\left(\widehat{\mathbf{X}}^{(k)}\right), \qquad \mathbf{Z}_{\mathrm{ctx}}^{(k+1)} = \mathcal{E}_{\mathrm{VAE}}\left(\mathbf{C}_k\right). \tag{5}$$

Here, Tail_r extracts the last r decoded frames to form the pixel-space context $\mathbf{C}_k$, and the VAE encoder $\mathcal{E}_{\mathrm{VAE}}$ maps it to the latent context $\mathbf{Z}_{\mathrm{ctx}}^{(k+1)}$. The first chunk uses a supplied motion frame $\mathbf{I}_{\mathrm{init}}$, giving a unified anchor definition for initialization and continuation:

$$\mathbf{Z}_{\mathrm{anc}}^{(k)} = \begin{cases} \mathcal{E}_{\mathrm{VAE}}(\mathbf{I}_{\mathrm{init}}), & k = 1, \\ \mathbf{Z}_{\mathrm{ctx}}^{(k)}, & k > 1. \end{cases} \tag{6}$$

The anchor remains clean throughout denoising. The reference image $\mathbf{I}$ and initial motion frame $\mathbf{I}_{\mathrm{init}}$ serve separate conditioning paths and need not coincide. Here, r counts decoded frames, whereas L_{anc} counts anchor positions in the compressed latent sequence. The decoded chunk $\widehat{\mathbf{X}}^{(k)}$ includes the context overlap, which is removed when assembling the final video.

Alongside this visual context, each chunk receives temporally aligned speech conditions. Acoustic features, SAT, and SAEs share a global conversation timeline, and chunk k selects the same ordered window $\mathcal{W}_k$ of audio-aligned positions for all three. The acoustic neighborhood in Eq. 1 is drawn from the full audio, retaining context outside the current chunk when available.

3.4 Multi-Reference Visual Conditioning

A generation chunk may span multiple shots with different compositions and participants, which a single initial reference image may not fully describe. MVC therefore extends the backbone's image conditioning with two complementary inputs: one reference image for the visual context of each shot, and an ID board of person crops that supplies identity cues across shots, including for participants absent from the initial image.

For a video with N shots, we separately VAE-encode the ID board and each shot reference as $\mathbf{I}_{\mathrm{ID}}$ and $\mathbf{I}_n$, respectively, where $n = 1, \ldots, N$ indexes shots. We concatenate these image latents into the reference prefix $\mathbf{I}_{\mathrm{ref}} = [\mathbf{I}_{\mathrm{ID}}, \mathbf{I}_1, \ldots, \mathbf{I}_N]$. Each reference contributes one latent-time position, giving a prefix length $K = N + 1$.

To combine these references with TCP, we prepend K conditioning positions to each chunk. At diffusion time τ, the video portion consists of the clean anchor $\mathbf{Z}_{\mathrm{anc}}^{(k)}$ from Eq. 6 and the remaining video latents $\mathbf{X}_{\tau,\mathrm{rest}}^{(k)}$ to be denoised. We augment the backbone's original image condition $\mathbf{I}_{\mathrm{orig}}$ and binary availability mask $\mathbf{m}_{\mathrm{orig}}$ with the reference prefix and its mask. This gives

$$\widetilde{\mathbf{X}}_{\tau}^{(k)} = [\mathbf{0}_K \mid \mathbf{Z}_{\mathrm{anc}}^{(k)} \mid \mathbf{X}_{\tau,\mathrm{rest}}^{(k)}], \qquad \widetilde{\mathbf{I}}_{\mathrm{ref}} = [\mathbf{I}_{\mathrm{ref}} \mid \mathbf{I}_{\mathrm{orig}}], \qquad \widetilde{\mathbf{m}}_{\mathrm{ref}} = [\mathbf{1}_K \mid \mathbf{m}_{\mathrm{orig}}]. \tag{7}$$

Here, $[\cdot \mid \cdot]$ concatenates along latent time, and a tilde denotes a prefix-augmented quantity. The zero-valued $\mathbf{0}_K$ reserves the prefix positions in the main latent, while the reference latents enter through the image-conditioning sequence. The binary mask marks available image conditions with one and unavailable positions with zero; its $\mathbf{1}_K$ prefix marks all references as available. The tensors $\mathbf{0}_K$ and $\mathbf{1}_K$ match the channel and spatial shapes of the main latent and availability mask, respectively. The reference positions provide fixed conditions during denoising and are removed before VAE decoding, so they do not become output video frames. Appendix A.4 details how the augmented sequences align with the audio conditions.

4 Experiments

4.1 Evaluation Setup

Datasets. We collect a legally authorized video dataset of about 2K hours for training, which covers the face or body of a single talking person and multiple-person conversations. **Single-person**: Following prior works (Kong et al., 2025; Yang et al., 2025), the main paper evaluates single-person animation on HDTF (Zhang et al., 2021) for talking heads and EMTD (Meng et al., 2024) for talking bodies. **Multi-person**: For multi-person conversational generation, we evaluate 30 conversation videos from MTHM (Kong et al., 2025), with at least 9 seconds of corresponding audio for each person shown in the reference image. We also construct MPMS, with 20 pairs of multi-person conversation covering turn-taking and off-screen speech. **Multi-shot**: We curate Movie Dialogues and derive evaluation subsets from AVA-AVD (Xu et al., 2022) and UnityShots (Huang et al., 2026) for audio-driven multi-shot generation.

Metrics. Sync-C and Sync-D (Chung & Zisserman, 2017) measure audio–lip synchronization; Expression-FID (E-FID), Fréchet Inception Distance (FID) (Heusel et al., 2017), and Fréchet Video Distance (FVD) (Unterthiner et al., 2018) measure expression, frame, and video quality, respectively. Higher Sync-C and lower Sync-D, E-FID, FID, and FVD are better. On MPMS, we additionally use TalkNet (Tao et al., 2021) scores above zero to detect visible speakers. Binding accuracy (BA) and Paired binding accuracy (PBA) require the predicted speaker set to exactly match SAT. Off-screen leakage rate (OLR) measures visible speaking detections during off-screen-only speech.

Implementation Details. We train Wan2.1-I2V-14B using AdamW on 32 NVIDIA H800-80G GPUs, with a learning rate of 5×10^{-5} after 100 warm-up steps. To establish audio-lip synchronization before learning multi-shot reference conditioning, we train in two stages. In Stage 1, we fine-tune the audio cross-attention layers and the audio adapter for SAM and TCP. In Stage 2, we train rank-32 LoRA adapters in the self-attention, cross-attention, and feed-forward layers of all DiT blocks for MVC. Appendix A details the datasets, MPMS protocol, sampling, and training.

4.2 Comparison with Competing Methods

Quantitative Evaluation. To verify the effectiveness of our method, we compare it with several state-of-the-art human animation method. As presented in Table 1, M^{2}Talk sets new state-of-the-art visual quality benchmarks on single-person generation, achieving the lowest E-FID and FID, representing a 9.6% and 4.8% FID reduction relative to InfiniteTalk. On multi-person conversational benchmarks (Table 2), our method achieved the highest Sync-C score on MTHM. Our method surpasses most other approaches across a majority of metrics, exhibiting superior performance in lip synchronization and video quality, which underscores the effectiveness of our approach.

Crucially, on the challenging MPMS benchmark designed for speaker attribution, M^{2}Talk drastically outperforms MultiTalk and Bind-Your-Avatar. While AnyTalker-14B also maintains top-tier lip-sync metrics, its rigid spatial binding mechanism fundamentally restricts natural human interactions, leading to static listener poses and rigid gestures during long turns, as further validated in our qualitative and user study.

Qualitative Evaluation. Figure 4 visualizes long-video comparisons. Baseline methods bound by rigid spatial masks exhibit persistent, frozen head/hand postures (AnyTalker) or suffer from background distortions and static listener behaviors (MultiTalk). In contrast, M^{2}Talk liberates listener dynamics without sacrificing lip-sync precision, producing rich non-verbal feedback, expressive

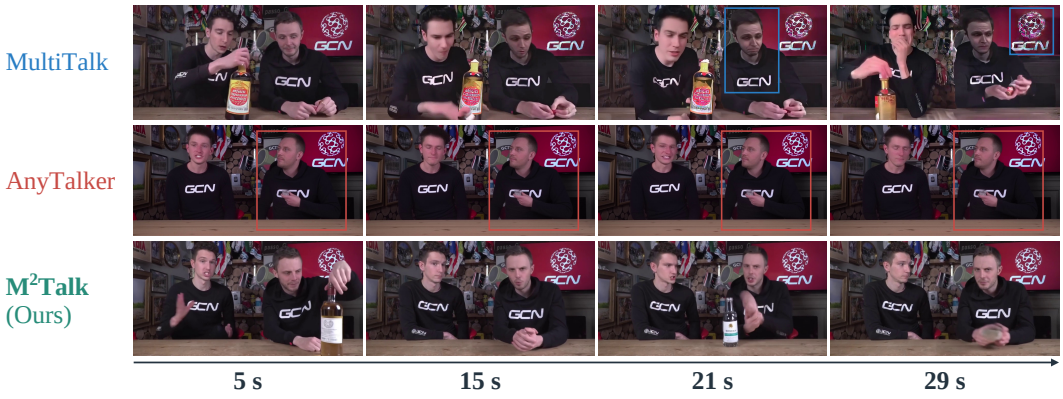


Figure 4: **Long-video comparison.** MultiTalk shows pronounced darkening of the right participant's face and background distortion at 29 s. AnyTalker's right participant retains nearly unchanged head and hand poses in the sampled frames from 5 to 29 s, illustrating limited motion progression over the sequence. M^{2}Talk sustains varied conversational gestures and object handling.

conversational gestures, and natural object handling. For this comparison, AnyTalker generates a single long sequence through its native inference interface.

Table 1: Quantitative comparison on single-person video generation. HDTF evaluates talking-head generation and EMTD evaluates talking-body generation. A dash (–) indicates that the model does not support the corresponding generation task.

Method	HDTF (talking head)					EMTD (talking body)				
	Sync-C↑	Sync-D↓	E-FID↓	FID↓	FVD↓	Sync-C↑	Sync-D↓	E-FID↓	FID↓	FVD↓
AniPortrait (Wei et al., 2024)	3.09	10.94	1.32	32.83	112.21	–	–	–	–	–
VExpress (Wang et al., 2024)	5.79	8.37	8.92	60.49	200.60	–	–	–	–	–
Hallo3 (Cui et al., 2025)	6.55	8.49	1.12	33.98	153.31	–	–	–	–	–
EchoMimic (Chen et al., 2025c)	5.36	8.99	1.27	60.82	240.07	–	–	–	–	–
Sonic (Ji et al., 2025)	8.35	**6.43**	1.22	29.53	**89.34**	–	–	–	–	–
EchoMimicV2 (Meng et al., 2024)	–	–	–	–	–	6.31	8.41	1.91	35.99	**163.60**
Fantasy Talking (Wang et al., 2025a)	3.61	10.78	1.36	32.64	103.01	3.32	11.41	1.98	37.68	284.29
Bind-Your-Avatar (Huang et al., 2025)	5.61	9.88	5.92	36.54	575.98	5.10	9.55	5.62	47.84	668.08
OmniAvatar (Gan et al., 2025)	7.06	8.63	1.23	26.63	112.49	6.93	8.55	1.38	29.47	308.14
HV-Avatar (Chen et al., 2025b)	7.08	8.44	1.35	29.73	330.80	6.35	9.08	1.57	43.14	431.02
MultiTalk (Kong et al., 2025)	8.53	6.81	1.24	27.27	124.06	8.34	7.30	1.51	31.93	238.77
AnyTalker-14B (Zhong et al., 2025)	8.85	7.25	1.27	23.77	226.26	8.31	7.47	1.47	31.43	322.01
InfiniteTalk (Yang et al., 2025)	**9.01**	6.97	0.84	14.42	133.56	**8.50**	**7.23**	1.18	28.12	282.45
InteractAvatar (Zhang et al., 2026)	8.31	7.63	1.16	24.46	351.48	7.81	7.76	1.49	35.40	425.03
M^{2}Talk (Ours)	8.65	7.45	**0.68**	**13.03**	127.40	8.38	7.52	**1.06**	**26.78**	280.03

Table 2: Quantitative comparison on multi-person video generation and user study.

Method	MTHM (Generation quality)					MPMS (Speaker binding)			User study
	Sync-C↑	Sync-D↓	E-FID↓	FID↓	FVD↓	BA↑	PBA↑	OLR↓	Mean↑
MultiTalk	8.11	7.41	**2.62**	98.27	693.57	63.95	55.07	0.00	6.83
Bind-Your-Avatar	5.55	9.88	4.85	77.64	697.20	56.32	39.39	0.00	–
AnyTalker-14B	8.41	**7.35**	2.64	**46.24**	**468.97**	**92.36**	**90.78**	0.00	1.83
M^{2}Talk (Ours)	**8.54**	7.46	2.70	89.81	695.91	91.11	88.22	0.00	**91.33**

User Study. M^{2}Talk leads all six dimensions (Figure 5). Qualitative breakdowns reveal that while spatially constrained baselines suffer from rigid postures during extended turns, our decoupling strategy empowers the generator to synthesize expressive, contextually aligned listener reactions (e.g., nods, gaze shifts, and posture adjustments).

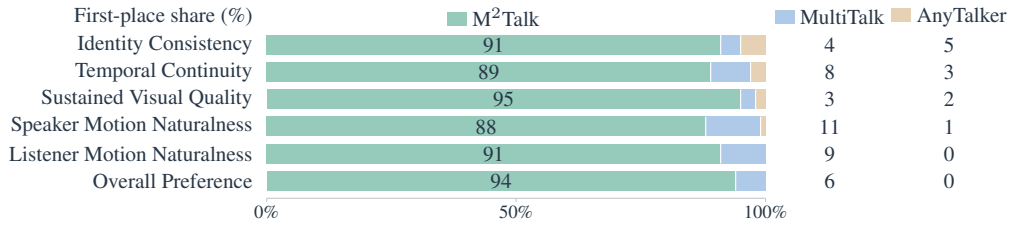


Figure 5: **User study preferences.** Appendix C gives study details and more comparisons.

4.3 Ablation Study

Table 3: **SAM and TCP ablations on MTHM and MPMS.** Subscripts show changes relative to the full model: improvement, degradation, or no change (black).

Method	MTHM (Generation quality)					MPMS (Speaker binding)		
	Sync-C↑	Sync-D↓	E-FID↓	FID↓	FVD↓	BA↑	PBA↑	OLR↓
w/o SAM	$8.48_{-0.06}$	$7.36_{-0.10}$	$2.63_{-0.07}$	$87.71_{-2.10}$	$693.50_{-2.41}$	$34.00_{-57.11}$	$25.77_{-62.45}$	$100.00_{+100.00}$
w/o TCP	$7.91_{-0.63}$	$7.94_{+0.48}$	$2.71_{+0.01}$	$90.40_{+0.59}$	$862.36_{+166.45}$	$88.36_{-2.75}$	$82.34_{-5.88}$	$0.00_{0.00}$
M²Talk (Ours)	8.54	7.46	2.70	89.81	695.91	91.11	88.22	0.00

Effect of SAM. With training data and steps fixed, SAM improves MPMS binding and eliminates off-screen leakage (Table 3). On MTHM, it slightly improves Sync-C, while Sync-D and visual metrics slightly worsen, indicating its primary benefit is speaker binding and off-screen handling.

Effect of TCP. TCP supports long-video continuation by carrying spatial appearance and temporal motion from recent generated frames into the next chunk. This context helps preserve ongoing facial and body movements and also improves lip synchronization (Table 3). Figure 10 illustrates its role: single-frame continuation loses the background motion direction, whereas TCP maintains it by propagating spatiotemporal information across chunk boundaries.

Multi-Shot Evaluation. We evaluate H3 (MiniMax, 2026), JoyAI-Echo 1.5 (Duan et al., 2026), TCR (Liu et al., 2026), and MoCha (Wei et al., 2025). Only audio-adapted H3 shares our model's use of multiple shot references and driving audio. Table 4 compares H3, M²Talk, and w/o MVC on Movie Dialogues. M²Talk leads in cut timing and exceeds H3 in reference-identity similarity and face coverage, while H3 scores higher in cross-shot consistency and frame quality.

Table 4: Ablation study of MVC and comparison with H3 (MiniMax, 2026) on Movie Dialogues.

Method	Cut timing					Identity and coverage				Consistency and quality			
	All↑	P↑	R↑	F1↑	MAE↓	ID↑	FC↑	X-ID↑	PC↑	CISC↑	BISC↑	AES↑	IQ↑
H3	6/40	33.33	17.78	23.19	76.0	0.673	64.8	**0.847**	42.9	**0.946**	0.811	0.530	**0.670**
M²Talk	**37/40**	**100.00**	**93.33**	**96.55**	**32.4**	**0.851**	**89.6**	0.823	**100.0**	0.877	0.662	0.527	0.642
w/o MVC	$0/40_{-37}$	$0.00_{-100.00}$	$0.00_{-93.33}$	$0.00_{-96.55}$	–	$0.579_{-0.272}$	$68.5_{-21.1}$	$0.805_{-0.018}$	$\mathbf{100.0}_{0.0}$	$0.860_{-0.017}$	$\mathbf{0.882}_{+0.220}$	$\mathbf{0.535}_{+0.008}$	$0.626_{-0.016}$

Effect of MVC. MVC improves shot-transition timing, reference-identity similarity, and face coverage (Table 4), with consistent trends on AVA-AVD and UnityShots. Without MVC, outputs can miss cuts or mix person appearances; the full model better preserves the intended composition. Appendix B provides dataset and input details, full results, and qualitative comparisons.

5 Conclusion

We presented M²Talk, a framework for multi-person and multi-shot conversational video generation driven by speaker-identity conditioning. By modulating acoustic features with Speaker-Aware Embeddings rather than imposing restrictive per-person spatial masks, M²Talk decouples speaker activity from visual localization, preserving natural speaking gestures and listener dynamics. Coupled with multi-reference conditioning and streaming temporal propagation, the framework maintains visual and cross-modal coherence across dynamic shot transitions and long-duration generation.

AI USE STATEMENT

We used generative AI tools to assist with manuscript preparation, literature review, and visualization. AI-assisted materials were reviewed and revised under author supervision. The authors take full responsibility for the accuracy, originality, and integrity of the final submission.

ETHICS STATEMENT

M^{2}Talk enables audio-driven multi-person multi-shot conversational video generation and shares the potential risks of video synthesis, including deceptive media, impersonation, and biases inherited from pretrained models and training data. Generated videos should be clearly identified as synthetic. The collection and use of real videos, as well as model deployment, should respect privacy, consent where applicable, and the licenses of source materials and pretrained models. The user-study protocol and aggregate preferences among generated videos are reported in Appendix C.1.

REPRODUCIBILITY STATEMENT

Appendix A details the datasets, sampling settings, reference and audio alignment, and training objective. Appendix B specifies multi-shot inputs, evaluation metrics, and complete results, while Appendix C provides the user-study protocol, TCP ablation, and additional qualitative comparisons. Further visual comparisons are available on our anonymous project page. We will publicly release the model weights.

REFERENCES

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL https://arxiv.org/abs/2006.11477.

Jiaben Chen, Zixin Wang, Ailing Zeng, Yang Fu, Xueyang Yu, Siyuan Cen, Julian Tanke, Yihang Chen, Koichi Saito, Yuki Mitsufuji, and Chuang Gan. TalkCuts: A large-scale dataset for multi-shot human speech video generation. *arXiv preprint arXiv:2510.07249*, 2025a. URL https://arxiv.org/abs/2510.07249.

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. *arXiv preprint arXiv:2505.20156*, 2025b.

Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *AAAI*, volume 39, pp. 2403–2410, 2025c.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian Conference on Computer Vision 2016 Workshops*, pp. 251–263, 2017.

Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21086–21095, 2025.

Nan Duan, Haoyang Huang, Weiyang Jin, Haoran Li, Yaowei Li, Yuming Li, Yijun Liu, Xin Lu, Xiaoxiao Ma, Yanwen Ma, Yaofeng Su, Yilang Sun, Haoyu Wang, Zeyue Xue, Songchun Zhang, and Junhao Zhuang. Long-horizon audio-visual generation for persistent stories and interactive worlds. *arXiv preprint arXiv:2608.23383*, 2026. URL https://arxiv.org/abs/2608.23383.

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, Ke Sun, Linrui Tian, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-s2v: Audio-driven cinematic video generation, 2025. URL https://arxiv.org/abs/2508.18621.

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.

Jiehui Huang, Yuechen Zhang, Bin Xia, Jiahao Wang, Xu He, Zhenchao Tang, Meng Chu, Xin Tao, Pengfei Wan, and Jiaya Jia. Unityshots: Memory-driven multi-shot audio-video generation with boundary-aware gating. *arXiv preprint arXiv:2606.21661*, 2026.

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.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, et al. Sonic: Shifting focus to global audio perception in portrait animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 193–203, 2025.

Jianwen Jiang, Weihong Zeng, Zerong Zheng, Jiaqi Yang, Chao Liang, Wang Liao, Han Liang, Yuan Zhang, and Mingyuan Gao. Omnihuman-1.5: Instilling an active mind in avatars via cognitive simulation. *arXiv preprint arXiv:2508.19209*, 2025.

Ozgur Kara, Krishna Kumar Singh, Feng Liu, Duygu Ceylan, James M. Rehg, and Tobias Hinz. Shotadapter: Text-to-multi-shot video generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *ICCV*, 2025.

Yichen Liu, Quanwei Zhang, Haozhe Wang, Donghao Zhou, Jiankun Zhang, Xiaojie Li, Yang Shi, Jiaming Liu, Ruihua Huang, Yingtian Zou, and Daquan Zhou. The missing temporal link: Temporal context routing for script-driven audio-video generation. *arXiv preprint arXiv:2609.02367*, 2026. URL https://arxiv.org/abs/2609.02367.

Xiangyang Luo, Qingyu Li, Xiaokun Liu, Wenyu Qin, Miao Yang, Meng Wang, Pengfei Wan, Di Zhang, Kun Gai, and Shao-Lun Huang. Filmweaver: Weaving consistent multi-shot videos with cache-guided autoregressive diffusion. *arXiv preprint arXiv:2512.11274*, 2025.

Xingpei Ma, Shenneng Huang, Jiaran Cai, Yuansheng Guan, Shen Zheng, Hanfeng Zhao, Qiang Zhang, and Shunsi Zhang. Playmate2: Training-free multi-character audio-driven animation via diffusion transformer with reward feedback. *arXiv preprint arXiv:2510.12089*, 2025.

Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. *arXiv preprint arXiv:2411.10061*, 2024.

MiniMax. Open general intelligence: MiniMax H3 is now open source, August 2026. URL https://www.minimax.io/news/minimax-h3-open-source. Official release announcement, August 3, 2026.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM International Conference on Multimedia*, pp. 484–492, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Tomáš Souček and Jakub Lokoč. TransNet V2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. URL https://arxiv.org/abs/2008.04838.

Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3927–3935, 2021.

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 *ECCV*, pp. 244–260. Springer, 2024.

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, 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, 2025. URL https://arxiv.org/abs/2503.20314.

Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024.

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 preprint arXiv:2504.04842*, 2025a.

Qinghe Wang, Xiaoyu Shi, Baolu Li, Weikang Bian, Quande Liu, Huchuan Lu, Xintao Wang, Pengfei Wan, Kun Gai, and Xu Jia. MultiShotMaster: A controllable multi-shot video generation framework. *arXiv preprint arXiv:2512.03041*, 2025b. URL https://arxiv.org/abs/2512.03041.

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

Cong Wei, Bo Sun, Haoyu Ma, Ji Hou, Felix Juefei-Xu, Zecheng He, Xiaoliang Dai, Luxin Zhang, Kunpeng Li, Tingbo Hou, Animesh Sinha, Peter Vajda, and Wenhu Chen. MoCha: Towards movie-grade talking character synthesis. *arXiv preprint arXiv:2503.23307*, 2025. URL https://arxiv.org/abs/2503.23307.

Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024.

Eric Zhongcong Xu, Zeyang Song, Satoshi Tsutsui, Chao Feng, Mang Ye, and Mike Zheng Shou. AVA-AVD: Audio-visual speaker diarization in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3838–3847, 2022. URL https://arxiv.org/abs/2111.14448.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, and Xiaoming Wei. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. *ArXiv*, abs/2508.14033, 2025. URL https://api.semanticscholar.org/CorpusID:280686279.

Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. StoryMem: Multi-shot long video storytelling with memory. *arXiv preprint arXiv:2512.19539*, 2025a. URL https://arxiv.org/abs/2512.19539.

Peixuan Zhang, Zijian Jia, Kaiqi Liu, Shuchen Weng, Si Li, and Boxin Shi. STAGE: Storyboard-anchored generation for cinematic multi-shot narrative. *arXiv preprint arXiv:2512.12372*, 2025b. URL https://arxiv.org/abs/2512.12372.

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 *CVPR*, 2023.

Youliang Zhang, Zhengguang Zhou, Zhentao Yu, Ziyao Huang, Teng Hu, Sen Liang, Guozhen Zhang, Ziqiao Peng, Shunkai Li, Yi Chen, Zixiang Zhou, Yuan Zhou, Qinglin Lu, and Xiu Li. Making avatars interact: Towards text-driven human-object interaction for controllable talking avatars. *arXiv preprint arXiv:2602.01538*, 2026. URL https://arxiv.org/abs/2602.01538.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3661–3670, 2021.

Zhizhou Zhong, Yicheng Ji, Zhe Kong, Yiying Liu, Jiarui Wang, Jiasun Feng, Lupeng Liu, Xiangyi Wang, Yanjia Li, Yuqing She, Ying Qin, Huan Li, Shuiyang Mao, Wei Liu, and Wenhan Luo. Anytalker: Scaling multi-person talking video generation with interactivity refinement. *arXiv preprint arXiv:2511.23475*, 2025. URL https://arxiv.org/abs/2511.23475.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: Speaker-aware talking-head animation. *ACM Transactions On Graphics*, 39(6): 1–15, 2020.

APPENDIX CONTENTS

A Dataset and Implementation Details **15**
A.1 Dataset Details . . . 15
A.2 Sampling Details . . . 15
A.3 MPMS Evaluation Protocol . . . 15
A.4 Reference and Audio Alignment . . . 16
A.5 Denoising and Training Objective . . . 17

B Multi-Shot Evaluation **18**
B.1 Benchmarks and Actual Inputs . . . 18
B.2 Metric Definitions and Coverage . . . 18
B.3 Complete Quantitative Results . . . 20
B.4 Qualitative Comparisons . . . 22

C Long Video Generation **25**
C.1 User Study . . . 25
C.2 TCP Ablation . . . 26
C.3 Additional Qualitative Comparisons . . . 27

D Societal Impacts **29**

E Computational Cost **30**

F Limitations and Future Work **30**
F.1 Limitations . . . 30
F.2 Future Work . . . 30

A Dataset and Implementation Details

A.1 Dataset Details

Our initial evaluation covers four benchmarks for audio-driven generation and speaker binding. HDTF (Zhang et al., 2021) evaluates talking-head animation, while EMTD (Meng et al., 2024) evaluates talking-body animation. Our MTHM (Kong et al., 2025) evaluation uses 30 conversational videos, with at least 9 seconds of corresponding audio for each participant shown in the reference image. These benchmarks measure audio–lip synchronization, facial expression, and visual quality using the metrics in Section 4.1. The three additional multi-shot evaluation sets are described in Appendix B.1.

MPMS. Our speaker-binding benchmark contains 20 pairs. Within each pair, the reference image, audio, and text prompt are shared; the correspondence between the two visible-person activity channels and the speakers is exchanged, while off-screen activity remains unchanged. The annotations cover single-speaker turns, brief overlaps, intervals without an active visible speaker, and off-screen speech. Person identities are assigned from the reference image and retained throughout each clip. Figure 6 shows MTHM examples and the reference image for MPMS case 10. The BA, PBA, and OLR protocols are detailed in Appendix A.3.



MTHM MTHM MPMS

Figure 6: **Dataset examples.** Two MTHM examples from Kong et al. (2025) (left and middle) and the reference image of our MPMS case 10 (right).

A.2 Sampling Details

Generation settings. For M^{2}Talk benchmark inference, we use 40 sampling steps, a timestep shift of 7, and text and audio guidance scales of 5. Outputs are generated at 25 fps, with reference images resized and center-cropped to the selected aspect-ratio bucket. Audio is resampled to 16 kHz mono for feature extraction.

Chunk-wise synthesis. For single-person evaluation and the 30-second comparison, each generation window contains 81 frames. The first window uses one frame as its motion anchor; each subsequent window reuses the previous nine generated frames and contributes 72 new frames. We remove overlapping context frames when stitching the windows and trim the result to the audio duration. The SAT remains aligned with the original audio timeline.

Negative prompts. The single-person and long-video inference pipeline uses:

> color gradients, color shifts, camera shake, bright tones, overexposed, static, blurred details, subtitles, paintings, worst quality, low quality, JPEG compression residue, ugly, incomplete, extra fingers, poorly drawn hands, poorly drawn faces, deformed, disfigured, misshapen limbs, fused fingers, still picture, messy background, walking backwards

The MPMS inference runner uses an expanded default that additionally includes *style, works, images, overall gray, three legs, many people in the background*, with *static* repeated after *images*.

A.3 MPMS Evaluation Protocol

Evaluation window and predictions. We first compute a score for each video or original/swapped pair, then average with equal weight; we do not pool frames across videos. Let N_{vid} and N_{pair} denote the numbers of evaluated videos and pairs, respectively. For video i at evaluated frame t, let

$G_{i,t} \subseteq \{\text{left}, \text{right}\}$ be the SAT-annotated set of visible speakers. The empty set denotes no visible speaker, and both people may be active simultaneously. TalkNet (Tao et al., 2021) predicts

$$\widehat{G}_{i,t} = \{p : \text{TalkNetScore}_{i,t,p} > 0\}, \qquad C_{i,t} = D_{i,t}\,\mathbf{1}[\widehat{G}_{i,t} = G_{i,t}], \tag{8}$$

where p indexes the two people, only valid predictions enter $\widehat{G}_{i,t}$, $D_{i,t} = 1$ if both predictions are valid (and zero otherwise), and $\mathbf{1}[\cdot]$ is the indicator function. Thus $C_{i,t}$ is one only for an exact, fully observed set match.

Binding accuracy (BA). We evaluate frames with at least one SAT-annotated visible speaker:

$$V_i = \{t : |G_{i,t}| > 0\}, \qquad \text{BA}_i = \frac{\sum_{t \in V_i} C_{i,t}}{|V_i|}, \qquad \text{BA} = \frac{1}{N_{\text{vid}}} \sum_{i=1}^{N_{\text{vid}}} \text{BA}_i. \tag{9}$$

Predicting both people when only the left person should speak is incorrect, as is predicting one person during annotated overlapping speech. Missing predictions remain in the denominator and contribute zero correctness.

Paired binding accuracy (PBA). Pair j contains an original sample u_j and a sample v_j generated after swapping the correspondence between the two visible-person activity channels and the speakers. Each sample is evaluated against its own SAT. We retain the original sample's single-visible-speaker frames, which are time-aligned with the swapped sample:

$$S_j = \{t : |G_{u_j,t}| = 1\}, \qquad \text{PBA}_j = \frac{\sum_{t \in S_j} C_{u_j,t} C_{v_j,t}}{|S_j|}, \qquad \text{PBA} = \frac{1}{N_{\text{pair}}} \sum_{j=1}^{N_{\text{pair}}} \text{PBA}_j. \tag{10}$$

Both samples must be correct at the same frame to receive credit. PBA is not the average of two BA scores. Missing predictions in either sample count as incorrect without removing the frame from the denominator.

Off-screen leakage rate (OLR). Let $o_{i,t} \in \{0, 1\}$ indicate whether the SAT off-screen channel is active. We evaluate only off-screen speech with no annotated visible speaker:

$$O_i = \{t : o_{i,t} = 1,\ G_{i,t} = \emptyset\}, \qquad \mathcal{I}_O = \{i : |O_i| > 0\}. \tag{11}$$

Any detected visible speaker constitutes leakage. With complete predictions,

$$\text{OLR}_i = \frac{\sum_{t \in O_i} \mathbf{1}[\widehat{G}_{i,t} \neq \emptyset]}{|O_i|}, \qquad \text{OLR} = \frac{1}{|\mathcal{I}_O|} \sum_{i \in \mathcal{I}_O} \text{OLR}_i. \tag{12}$$

Videos without eligible off-screen frames are excluded from this average. For frames with missing predictions, we compute a lower bound treating these frames as non-leaking and an upper bound treating them as potentially leaking. Specifically, replace the per-frame indicator by $\ell^-_{i,t} = D_{i,t}\,\mathbf{1}[\widehat{G}_{i,t} \neq \emptyset]$ for the lower bound and by $\ell^+_{i,t} = \max(\ell^-_{i,t}, 1 - D_{i,t})$ for the upper bound, retaining the same denominators and macro-average. The two bounds coincide for all results in Tables 2 and 3. All scores are reported as percentages.

A.4 Reference and Audio Alignment

Reference shapes and masks. If the current chunk's video portion has L_Z latent positions, including L_{anc} anchor positions, all three augmented sequences have $K + L_Z$ positions and matching spatial sizes. These lengths may differ between initialization and continuation. The tensors $\mathbf{0}_K$ and $\mathbf{1}_K$ have K positions and the channel/spatial shapes required by the main latent and reference mask, respectively.

The subscript orig denotes the backbone's original image-conditioning path. Its condition $\mathbf{I}_{\text{orig}}$ is VAE-encoded from $\mathbf{I}$ followed by zero-valued frames. The condition and mask match the current chunk; their chunk indices, and those of the augmented image tensors, are omitted for readability. Let j index frame positions in this conditioning sequence before adding the prefix. The mask $\mathbf{m}_{\text{orig}}$ marks the first position ($j = 0$), with one indicating a supplied image and zero indicating no supplied image. It is packed onto the latent-time grid before concatenation. For multi-shot clips, $\mathbf{I}$ is the

selected shot 1 reference image; it need not be the video's first frame. Reference-prefix mask entries are also one. These masks describe image conditioning, not speaker activity or a training loss mask.

Aligned Audio Conditions. The acoustic windows in Eq. 1 repeat the nearest boundary feature for indices outside the full audio sequence. For chunk k, let $\mathcal{W}_k$ be its ordered window of global audio-aligned positions. SAM supplies the final audio condition

$$\mathbf{a}_{\text{cond}}^{(k)} = \Phi_{\text{SAM}}(\mathbf{a}_{\mathcal{W}_k}, \mathbf{M}_{\mathcal{W}_k}; K)\,. \tag{13}$$

The operator Φ_{SAM} comprises reference-prefix alignment, SAE construction, FiLM, and the audio adapter $\mathcal{P}_a$. Before FiLM, acoustic blocks and SAT receive silent prefix entries and repeated first-position entries to preserve the VAE's causal temporal grouping. The SAE and FiLM equations also apply to these aligned sequences. After modulation, $\mathcal{P}_a$ handles the first augmented position separately, groups the remaining positions according to the VAE temporal compression, and applies learned projections. Its output aligns with all $K + L_Z$ latent positions and supplies the keys and values of frame-aligned audio cross-attention.

A.5 Denoising and Training Objective

Denoising with All Conditions. At each denoising step, the DiT predictor f_θ, with learned parameters θ, receives

$$f_\theta\Big(\,\text{Concat}_{\text{ch}}\left(\widetilde{\mathbf{X}}_\tau^{(k)}, \widetilde{\mathbf{m}}_{\text{ref}}, \widetilde{\mathbf{I}}_{\text{ref}}\right), \tau; \\ \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{image}}, \mathbf{a}_{\text{cond}}^{(k)}\Big). \tag{14}$$

Here, $\text{Concat}_{\text{ch}}$ denotes channel-wise concatenation, and f_θ includes patch embedding. The sampler uses its output to update the noisy video latents, keeps $\mathbf{Z}_{\text{anc}}^{(k)}$ clean, and removes the K reference positions before VAE decoding. In attention diagrams, $\mathbf{H}_v$ denotes visual token features inside the DiT; $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote projected queries, keys, and values. Bold $\mathbf{K}$ is distinct from the scalar prefix length K.

Training objective. We use rectified-flow training with the backbone's velocity parameterization. Let $\mathbf{X}_{0,\text{rest}}^{(k)}$ be the clean video latents outside the reference prefix and anchor, and let $\boldsymbol{\epsilon}$ be standard Gaussian noise of the same shape. At normalized noise level $\tau \in [0, 1]$, the noisy input and target velocity are

$$\mathbf{X}_{\tau,\text{rest}}^{(k)} = (1-\tau)\mathbf{X}_{0,\text{rest}}^{(k)} + \tau\boldsymbol{\epsilon}, \qquad \mathbf{v}^\star = \boldsymbol{\epsilon} - \mathbf{X}_{0,\text{rest}}^{(k)}. \tag{15}$$

The normalized noise level is obtained by applying the scheduler's timestep transformation to a uniform sample. Let $\widehat{\mathbf{v}}_{\theta,\text{rest}}^{(k)}$ denote the output of Eq. 14 restricted to the same video positions, and let Ω_k index their valid scalar latent entries after applying the data-provided ignore mask. The per-sample loss is

$$\ell_k(\theta) = \frac{1}{\max(|\Omega_k|, 1)} \sum_{u \in \Omega_k} \left(\widehat{v}_{\theta,\text{rest},u}^{(k)} - v_u^\star\right)^2. \tag{16}$$

The ignore mask is broadcast across latent channels. The reference prefix and clean anchors are held fixed and excluded from the prediction and target before evaluating the loss. The training loop averages the per-sample losses within each batch and applies the configured data-bucket weight. This loss mask is separate from the image-conditioning mask $\widetilde{\mathbf{m}}_{\text{ref}}$ and the SAT.

B Multi-Shot Evaluation

B.1 Benchmarks and Actual Inputs

Evaluation sets. We use AVA-AVD and UnityShots to denote our derived evaluation subsets, and Movie Dialogues to denote our curated movie-dialogue set. The AVA-AVD subset contains 40 source-selected windows from 15 movies in the official AVA-AVD test split (Xu et al., 2022), using the original movie speech: 38 windows request two shots and two request three. The UnityShots subset contains 12 public synthetic reference stories (eight continuous and four alternating) derived from UnityShots (Huang et al., 2026). Movie Dialogues contains 40 windows from 13 source movies, with original speech: 35 request two shots and five request three.

Table 5: **Actual multi-shot inference inputs.** Every configuration receives timestamped shot text. FF: independent first-frame condition; SR: ordered shot references; audio: shared driving speech; SAT: supplied speaker activity; IDB: independent ID board. H3's fixed-latent audio adapter is specific to this evaluation.

Method	FF	SR	Audio	SAT	IDB
H3	–	All	Fixed latent	–	–
JoyAI-Echo 1.5	Yes	All	–	–	–
TCR	Yes	–	–	–	–
MoCha	Yes	–	Yes	–	–
w/o MVC	Yes	All	Yes	Yes	–
M^2Talk	Yes	All	Yes	Yes	Yes

Baseline interfaces. H3 (MiniMax, 2026) receives ordered shot references and holds the source-audio latent fixed at each sampling step, without SAT or an independent motion-anchor branch. JoyAI-Echo 1.5 (Duan et al., 2026) receives the first frame and shot references through memory image slots, with silent memory audio. MoCha (Wei et al., 2025) uses the official demo's timestamped text, first-frame, and shared-speech interface, without later-shot references, SAT, or an ID board. TCR (Liu et al., 2026) receives a timestamped shot script and the first frame, without later-shot reference images. The latter two generate their own audio. The first frame differs from the first shot reference in all AVA-AVD and Movie Dialogues samples, but they are identical in all UnityShots samples; H3 therefore receives that same image through its first shot reference on UnityShots. Table 5 describes the interfaces exercised here, rather than each model's full capabilities.

B.2 Metric Definitions and Coverage

Cut timing. The frozen TransNet V2 detector (Souček & Lokoč, 2020) uses a threshold above 0.5. Contiguous transition detections are collapsed to their midpoint, with the pinned frame-end timestamp convention $(\text{midpoint} + 1)/\text{fps}$. We match detected and exact target boundaries one to one within ± 0.20 s, first maximizing match count and then minimizing total absolute error. Precision, recall, and F1 pool boundary counts across the dataset. Precision is undefined when no cut is detected; recall and F1 are zero when all requested cuts are missed. A clip is fully correct only when every target is matched and no extra cut is detected. MAE averages matched boundaries only; it must be read with recall. This detector measures transition timing, not shot semantics, lip synchronization, or speaker assignment. Every sample requests a cut, so no-cut false-positive rates are not tested.

Identity and coverage. ID and X-ID use InsightFace buffalo_l face embeddings with a 0.5 detection threshold, 640-pixel detection size, and 5 Hz sampling. ID measures generated-to-reference face cosine similarity; X-ID compares the same annotated identity across distinct requested shots within one generated clip. Source-derived face regions are frozen before scoring, transformed through each method's image geometry, and expanded by 1.25; ambiguous or missing generated faces reduce coverage. Grid points whose native-frame rounding crosses a target cut are excluded for all methods. Similarities average valid observations, without imputing zero for missing faces. Scores and face coverage first aggregate within each source movie/story and then weight eligible sources equally;

pair coverage is the fraction of frozen pairs with valid scores. Thus high similarity alone need not imply broad identity preservation.

The fixed face-opportunity counts are 677, 256, and 866 for AVA-AVD, UnityShots, and Movie Dialogues. Their X-ID denominators are respectively 9 pairs across 9 clips and 6 movies, 9 pairs across 8 stories, and 7 pairs across 5 clips and 3 movies. Movie Dialogues retains 83 source-portrait detection failures in the full 866-opportunity ID denominator; these are not generated-face failures. UnityShots uses static reference-image regions rather than temporal face tracks, and its published identity portraits also condition M^2Talk through the ID board. These region-based metrics assume the requested composition; reframing can reduce coverage even when a face appears elsewhere.

Cross-shot and frame quality. Character/background inter-shot consistency (CISC/BISC) uses raw ViCLIP cosine similarity from eight deterministic middle-bin frames per requested shot. CISC pairs share a source identity; BISC pairs use a separate source/reference-only background review, excluding uncertain pairs. The CISC/BISC pair counts are 16/23 on AVA-AVD, 8/4 on UnityShots, and 5/34 on Movie Dialogues. These are derived labels and implementation-specific scores, not official benchmark totals. Fixed target intervals can give high consistency when a method fails to cut; Table 6 must therefore accompany Table 7. AES and IQ are only the aesthetic-quality and imaging-quality submetrics of VBench (Huang et al., 2024), with the frozen official normalization and IQ's `longer` preprocessing. Visual scores aggregate within sources and then across sources, retaining undefined values as missing. No audio-quality or lip-sync scores are computed in this evaluation.

B.3 Complete Quantitative Results

Table 6: **Complete cut-timing results.** All is the number of clips with every requested cut matched and no extra cut. P/R/F1 are micro percentages at ± 0.20 s. MAE (ms) averages matched cuts only. A dash (–) denotes undefined P when no cut is detected, or undefined MAE when no cut is matched. † denotes self-generated audio, without the shared driving speech.

Method	All↑	P↑	R↑	F1↑	MAE↓
AVA-AVD (40 clips)					
H3	8/40	21.95	21.43	21.69	86.7
JoyAI-Echo 1.5†	5/40	14.63	14.29	14.46	60.0
TCR†	37/40	100.00	92.86	96.30	66.5
MoCha	0/40	0.00	0.00	0.00	–
w/o MVC	9/40	29.73	26.19	27.85	80.0
M²Talk	39/40	97.62	97.62	97.62	48.8
UnityShots (12 clips)					
H3	8/12	88.89	66.67	76.19	70.8
JoyAI-Echo 1.5†	2/12	12.50	16.67	14.29	60.0
TCR†	11/12	100.00	91.67	95.65	50.8
MoCha	0/12	0.00	0.00	0.00	–
w/o MVC	0/12	25.00	8.33	12.50	200.0
M²Talk	12/12	100.00	100.00	100.00	33.3
Movie Dialogues (40 clips)					
H3	6/40	33.33	17.78	23.19	76.0
JoyAI-Echo 1.5†	1/40	6.90	4.44	5.41	80.0
TCR†	2/40	25.00	13.33	17.39	126.1
MoCha	0/40	–	0.00	0.00	–
w/o MVC	0/40	0.00	0.00	0.00	–
M²Talk	37/40	100.00	93.33	96.55	32.4

Timing interpretation. M²Talk achieves cut F1 of 97.62%, 100.00%, and 96.55% on AVA-AVD, UnityShots, and Movie Dialogues. Movie Dialogues retains the original case with a one-frame middle shot at 1.60–1.64 s; prompt-time rounding and transition merging can affect this case. All Movie Dialogues rows retain the same 40-case denominator and exact frozen target times, without manual correction of detected cuts. The reported MAE is conditional on a successful match, rather than an error assigned to missed cuts.

MoCha's official demo has no matched target boundary on any of the three sets, yielding 0% recall/F1 and undefined MAE. It produces two unmatched detected boundaries on each of AVA-AVD and UnityShots, and none on Movie Dialogues; precision on Movie Dialogues is therefore also undefined.

Table 7: **Complete visual results.** ID and X-ID are raw face-embedding cosines; FC and PC are face and cross-shot pair coverage (%). CISC/BISC are raw ViCLIP cosines; AES/IQ are the two VBench submetrics. Higher is better, but consistency can reward missed cuts and must be read with Table 6. †: no shared driving speech.

Method	ID	FC	X-ID	PC	CISC	BISC	AES	IQ
AVA-AVD								
H3	0.741	91.3	0.572	100.0	0.739	0.695	0.519	0.588
JoyAI-Echo 1.5†	0.546	75.5	0.542	66.7	0.655	0.686	0.517	0.570
TCR†	0.364	85.3	0.334	66.7	0.577	0.524	0.533	0.614
MoCha	0.413	69.2	0.821	55.6	0.916	0.921	0.504	0.605
w/o MVC	0.566	74.7	0.398	77.8	0.659	0.715	0.526	0.567
M^{2}Talk	0.798	96.5	0.551	100.0	0.671	0.640	0.525	0.557
UnityShots								
H3	0.581	76.8	0.545	44.4	0.723	0.654	0.624	0.678
JoyAI-Echo 1.5†	0.410	85.0	0.416	77.8	0.671	0.692	0.614	0.671
TCR†	0.434	79.9	0.236	44.4	0.428	0.429	0.613	0.670
MoCha	0.372	71.5	0.143	44.4	0.900	0.850	0.511	0.668
w/o MVC	0.404	93.8	0.476	77.8	0.824	0.936	0.627	0.634
M^{2}Talk	0.548	100.0	0.469	100.0	0.501	0.594	0.632	0.671
Movie Dialogues								
H3	0.673	64.8	0.847	42.9	0.946	0.811	0.530	0.670
JoyAI-Echo 1.5†	0.492	63.2	0.671	57.1	0.821	0.845	0.537	0.674
TCR†	0.533	53.7	0.884	14.3	0.869	0.899	0.514	0.651
MoCha	0.437	55.9	0.910	28.6	0.979	0.955	0.503	0.624
w/o MVC	0.579	68.5	0.805	100.0	0.860	0.882	0.535	0.626
M^{2}Talk	0.851	89.6	0.823	100.0	0.877	0.662	0.527	0.642

Visual interpretation. M^{2}Talk improves ID and face coverage over w/o MVC on all three sets, but it does not lead every visual metric. H3 has higher X-ID on AVA-AVD and UnityShots, with only 44.4% pair coverage on the latter. TCR's X-ID of 0.884 on Movie Dialogues covers only one of seven planned pairs (14.3%). These conditional scores should not be read as dataset-wide identity success. High CISC/BISC from w/o MVC or other methods with low cut recall can reflect continued appearance within a shot rather than successful consistency across the requested transitions.

MoCha's X-ID of 0.910 on Movie Dialogues covers only two of seven planned pairs (28.6%), while ID is 0.437 at 55.9% face coverage. Its CISC/BISC of 0.979/0.955 coexist with zero detected cuts, illustrating why high consistency alone does not establish successful multi-shot generation.

B.4 Qualitative Comparisons

Figures 7–9 compare all six configurations on Movie Dialogues, AVA-AVD, and UnityShots at matched output timestamps. The Shot refs row illustrates the intended shot sequence using reference images; model-specific inputs are listed in Table 5. Frames retain their full field of view and native aspect ratios.

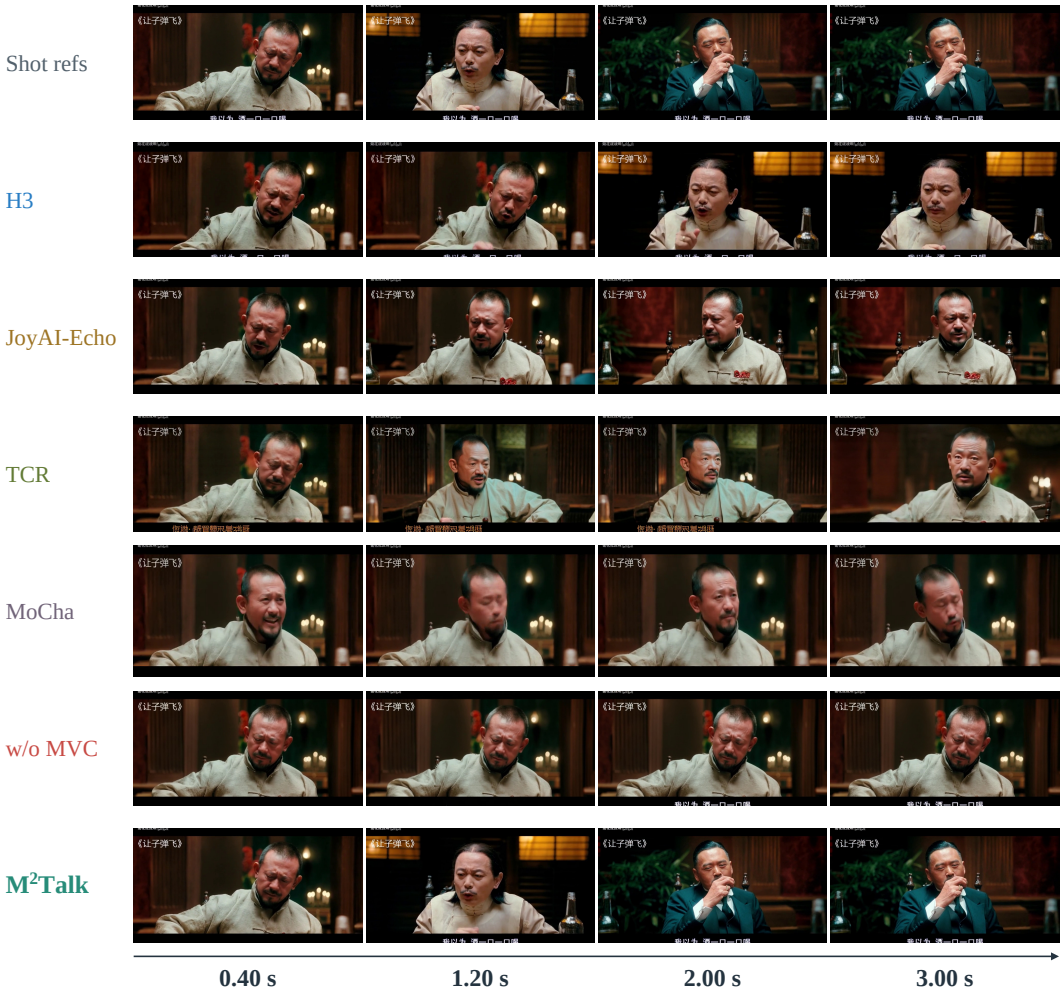


Figure 7: **Multi-shot comparison on Movie Dialogues.** A three-person conversation requests cuts at 0.68 and 1.72 s. Frames at matched timestamps show how the methods follow the requested character changes and framing.

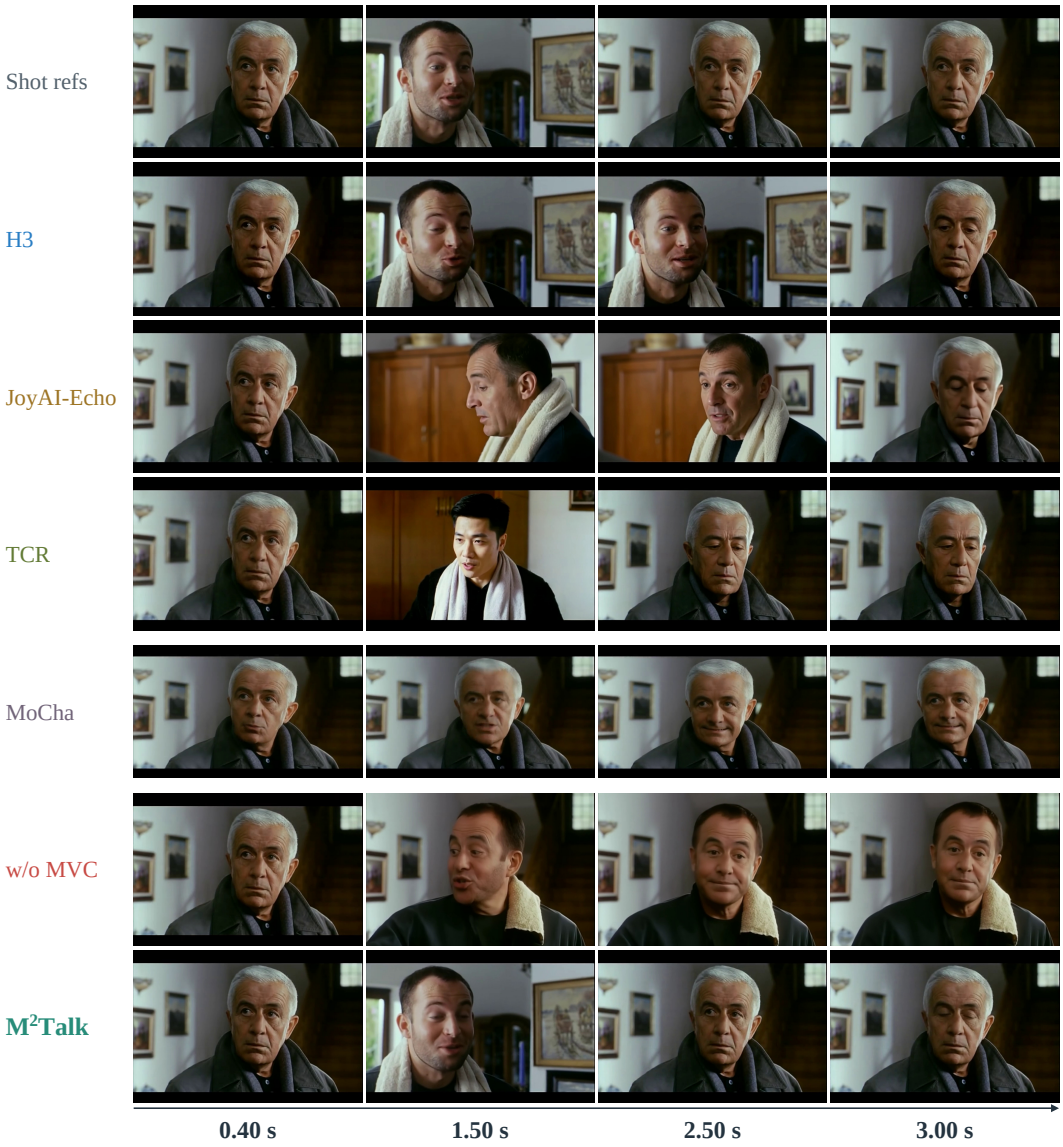


Figure 8: **Multi-shot comparison on AVA-AVD.** Requested cuts at 0.72 and 2.28 s switch to the second character and return to the first. Matched frames show the timing of the return shot and character appearance across views.

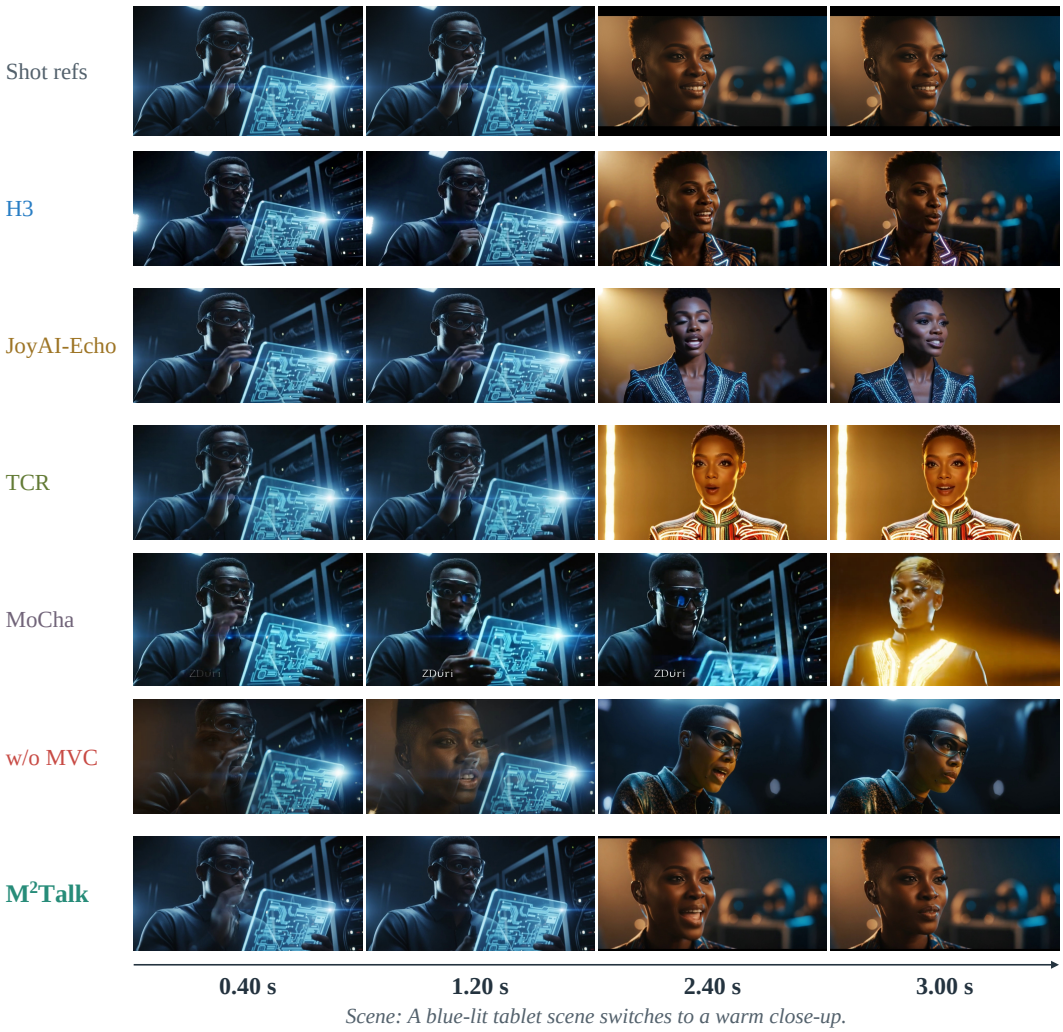


Figure 9: **Multi-shot comparison on UnityShots.** The requested cut at 1.60 s changes a blue-lit scene to a warm close-up. The comparison illustrates differences in framing and character appearance after the requested transition.

C LONG VIDEO GENERATION

C.1 USER STUDY

Protocol. Ten participants complete a web-based comparison of M^{2}Talk, MultiTalk, and AnyTalker on ten 30-second samples. All participants evaluate the same samples. Sample order and anonymous A/B/C method assignments are randomized per participant; each method occupies each letter position three or four times across the ten samples. The three videos share one audio track and can be played synchronously, paused, replayed, or enlarged. Within each dimension, participants give a strict best-to-worst ranking, with 1 best and 3 worst, or mark an item not applicable (N/A). No answers are preselected.

Judgment criteria. The questionnaire uses the following six dimensions:

- **Identity Consistency:** Do faces, hairstyles, and clothing remain consistent throughout the generated video, without identity changes or swaps?
- **Temporal Continuity:** Are movements continuous, without sudden jumps, pose resets, or visible chunk boundaries?
- **Sustained Visual Quality:** Does the video remain clear and intact, without increasing blur, color shifts, or facial, hand, and background artifacts?
- **Speaker Motion Naturalness:** Are the speaker's expressions and head/body movements natural and appropriate to the speech rhythm and emotion?
- **Listener Motion Naturalness:** Are the listener's gaze, expressions, and reactions natural and appropriate, without apparent speaking, rigidity, or mechanical repetition?
- **Overall Preference:** Which version provides the preferred overall viewing experience?

The reference image is not shown, so identity is judged within each generated video. Instructions distinguish camera-facing behavior in video calls or broadcasts from interaction with nearby people in natural scenes, and do not treat intended shot changes as discontinuities. Speaker/listener roles follow the conversation; during off-screen-only speech, visible people are treated as listeners. Viewing is self-paced, with no minimum coverage enforced.

Table 8: Long-video user study: 10 participants rank three methods on 10 30-second samples. Each dimension has 100 valid rankings. Mean rank (MR) is lower-is-better; first-place percentage (First, %) is higher-is-better. Best results are bold.

Dimension	MultiTalk		AnyTalker		M^{2}Talk (Ours)	
	MR↓	First↑	MR↓	First↑	MR↓	First↑
Identity Consistency	2.31	4.0	2.56	5.0	**1.13**	**91.0**
Temporal Continuity	2.23	8.0	2.64	3.0	**1.13**	**89.0**
Sustained Visual Quality	2.33	3.0	2.59	2.0	**1.08**	**95.0**
Speaker Motion Naturalness	2.04	11.0	2.82	1.0	**1.14**	**88.0**
Listener Motion Naturalness	2.04	9.0	2.85	0.0	**1.11**	**91.0**
Overall Preference	2.13	6.0	2.79	0.0	**1.08**	**94.0**

Aggregation and results. We include only the ten completed formal submissions, excluding preview and incomplete sessions. All 100 participant–sample comparisons contain six valid rankings, with no N/A items, yielding 600 ranking tasks. Table 8 reports mean rank and the fraction of rankings placing each method first, weighting every participant–sample comparison equally. The user-study column in Table 2 is the arithmetic mean of these six first-place percentages. M^{2}Talk leads all six dimensions; its first-place share is 88% for speaker motion, 91% for listener motion, and 94% overall. These are descriptive results from ten participants with repeated judgments, without a significance claim.

C.2 TCP Ablation

Figure 10 compares single-frame continuation with the full nine-frame TCP context at matched output timestamps. Under single-frame continuation, the background roller coaster reverses direction near 16 s; TCP maintains the original direction over the illustrated interval.



Figure 10: **TCP ablation on motion continuity.** Frames are sampled at matched output timestamps, with enlarged background regions below. Blue arrows indicate the original motion direction; orange arrows mark reversed motion in the ablation (w/o TCP).

C.3 Additional Qualitative Comparisons

Figures 11–13 show nine further 30-second examples, covering conversations, video calls, broadcast interviews, and animation. Appendix panels use 1, 10, 20, and 29 seconds; Figure 4 uses 5, 15, 21, and 29 seconds to highlight pose persistence across speaking and listening turns. Times are matched across methods, with complete frames at native aspect ratios and no image enhancement.

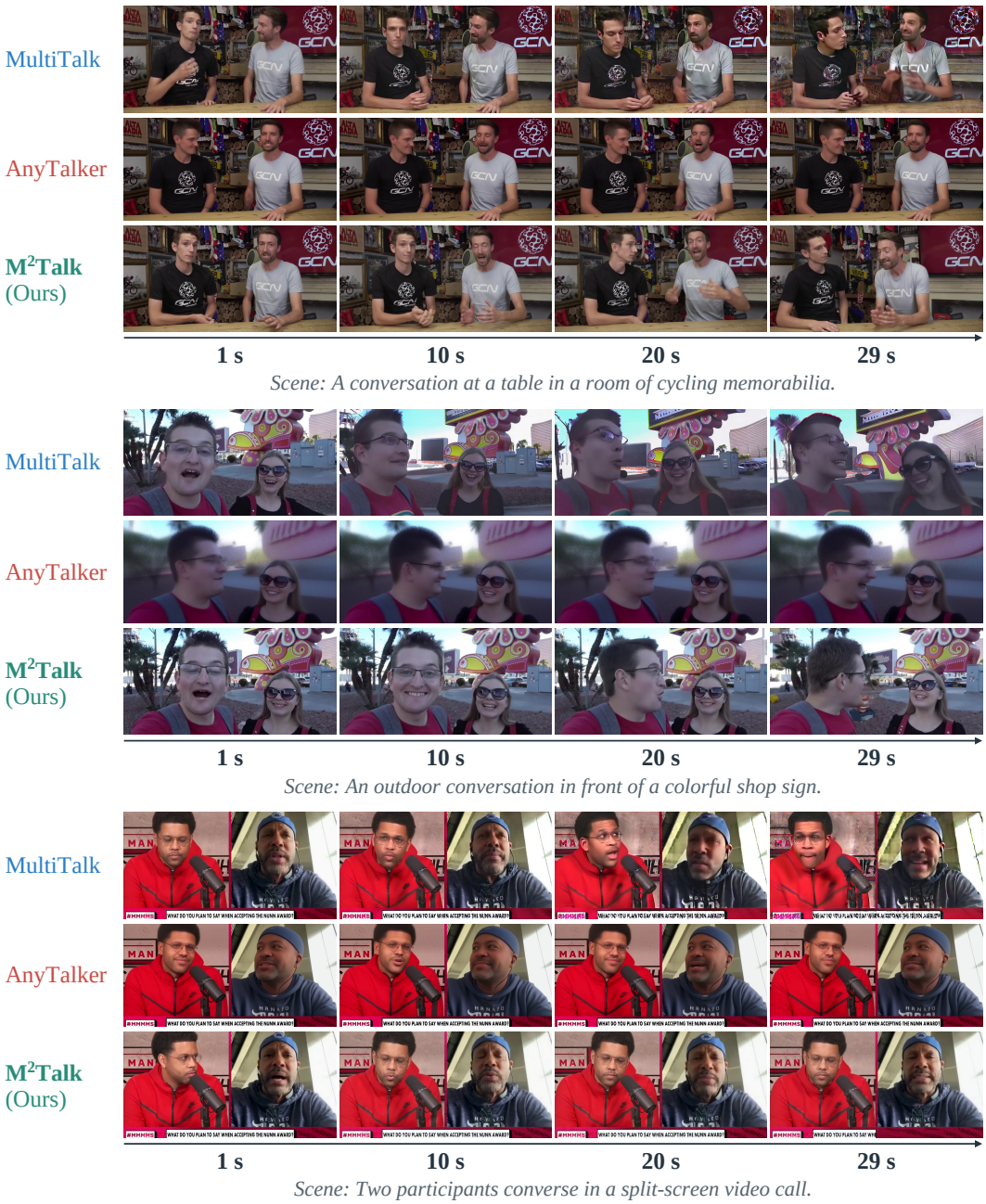


Figure 11: **Long-video comparisons: indoor/outdoor conversations and a video call.**

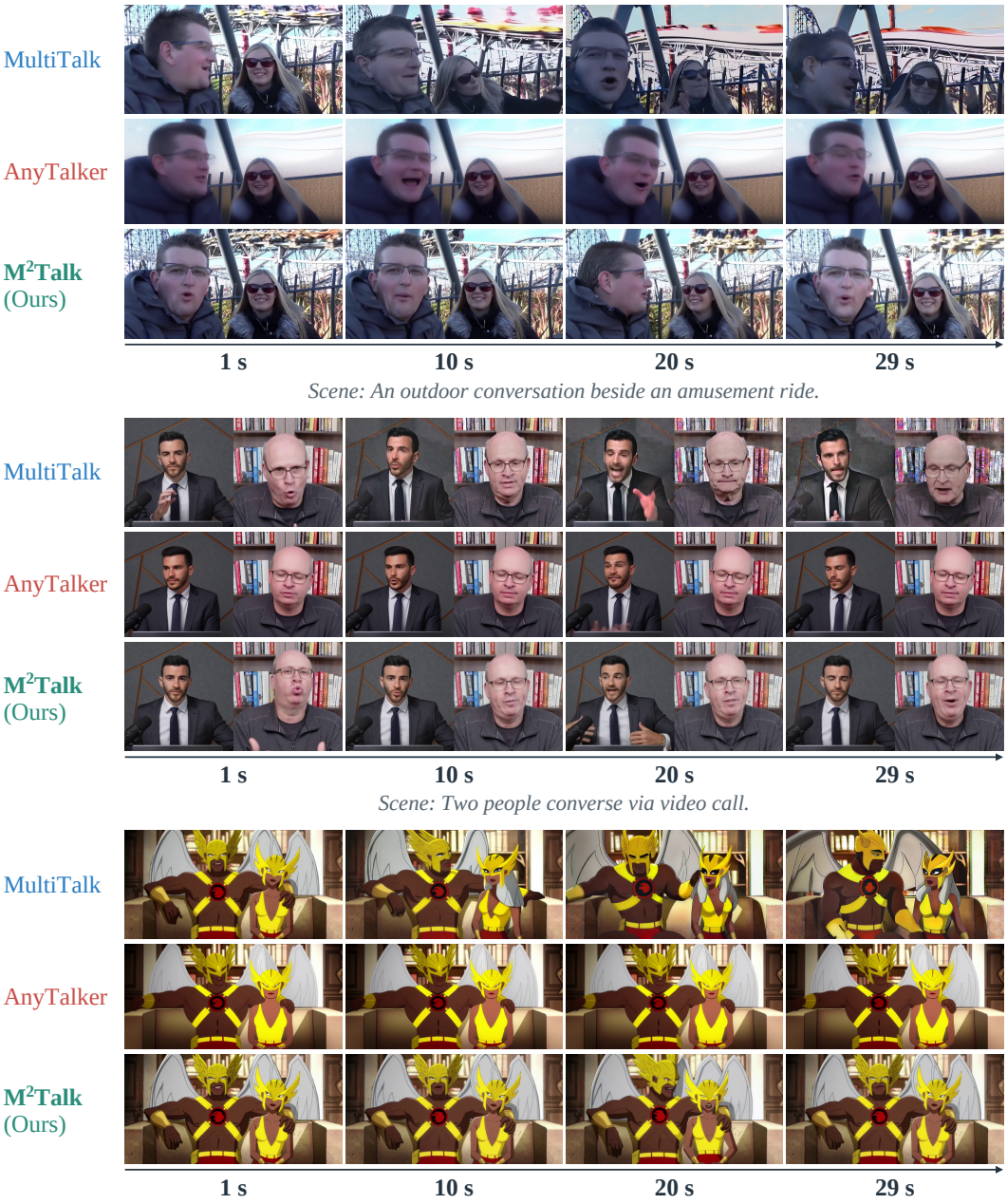
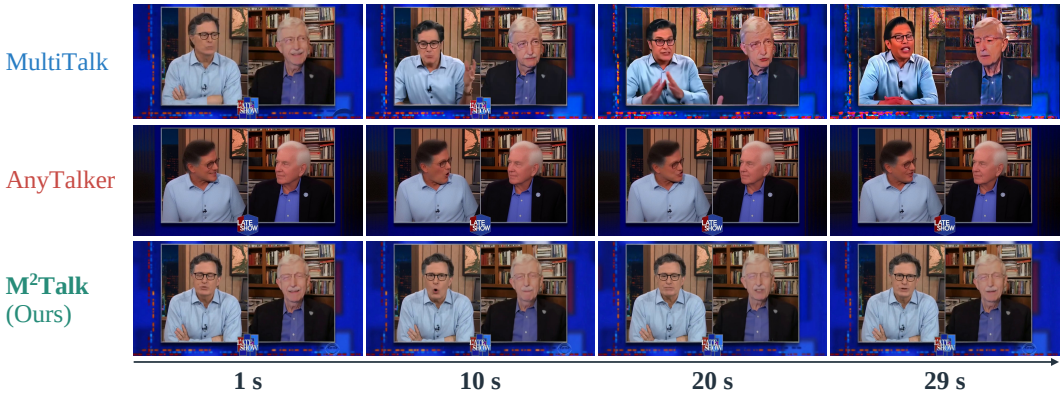


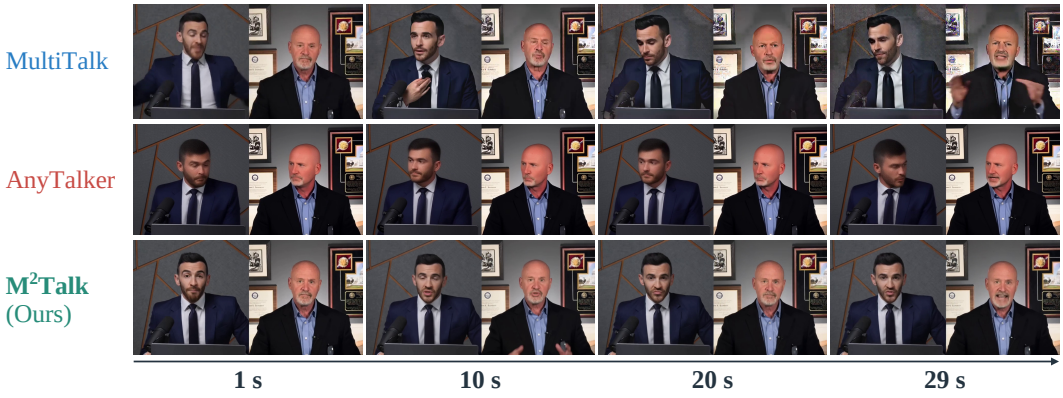
Figure 12: **Long-video comparisons: outdoor conversation, video call, and animation.**



Scene: A split-screen news interview with two presenters.



Scene: A television interview between a host and a guest.



Scene: A video conversation between two people in office settings.

Figure 13: **Long-video comparisons: broadcast interviews and a video call.**

D SOCIETAL IMPACTS

This paper introduces an effective tool for audio-driven multi-person multi-shot conversational video generation to the community. However, there exists a risk wherein malicious entities could exploit this framework to generate fake videos of celebrities, potentially misleading the public. This concern is not unique to our approach but is a shared consideration across various human animation methodologies.

E COMPUTATIONAL COST

Table 9 compares added parameters and mean denoising time per step for 81-frame windows. Timings use a single NVIDIA H100-80GB GPU with each method's native resolution and guidance settings. Parameter counts include all modules added to the Wan2.1 I2V backbone, excluding pretrained encoders.

Table 9: Additional parameters and per-step denoising time for 81-frame windows.

Method	Extra params. (B)	Resolution	Time/step (s)
MultiTalk	**2.487**	896×448	14.14
AnyTalker	4.593	864×416	**9.06**
M^{2}Talk	2.489	896×448	11.68

The shared audio adapter and cross-attention modules account for 2.487 B parameters. MultiTalk's L-RoPE adds no learned parameters; AnyTalker's face-conditioning modules add a further 2.106 B. Our speaker embedding and FiLM modulator add only 1.78 M, keeping the total close to MultiTalk while using fewer parameters than AnyTalker.

F LIMITATIONS AND FUTURE WORK

F.1 LIMITATIONS

The framework relies on a large video diffusion backbone and requires substantial computational resources. Higher-resolution or longer video generation further increases inference cost.

F.2 FUTURE WORK

Separating speaker activity from visual localization provides a control interface for further conversational generation tasks. Combining it with diarization and visual tracking could automatically update speaker activity and visibility for interactive conversations. Editing identity assignments and turn schedules could support controllable dialogue dubbing and off-screen narration while retaining visual references and ongoing motion. Persistent identity memory could extend the fixed speaker slots to larger casts and repeated appearances across long narratives. Combining this binding interface with distilled video generators and cached streaming inference is another direction, aiming to reduce sampling cost while maintaining speaker assignment and temporal continuity.