ANYTALKER: DATA-EFFICIENT SCALING OF MULTI-PERSON TALKING VIDEO GENERATION

Anonymous authors
Paper under double-blind review



Figure 1: AnyTalker generates natural multi-person talking videos rich in emotion, cross-speaker interactivity, and generalizes to varied identity counts and non-human characters. Videos with sound: anonymous demo page.

ABSTRACT

Generating multi-person talking videos requires modeling not only per-speaker lip synchronization but also cross-speaker interactions, including gaze shifts, nods, and listening reactions. Existing methods often rely on hundreds or thousands of hours of curated multi-person video, which is costly to annotate and concentrated in domains such as studio podcasts. In this paper, we introduce AnyTalker, a data-efficient framework that shows how to build multi-person generation largely from abundant, open-domain single-person data. To this end, we concatenate independent single-person clips to teach the model to bind each audio stream to its corresponding face. To extend this learned binding beyond the two-person training setting, each DiT block uses an Audio-Face Cross Attention module with parameters shared across identity-audio pairs, allowing the number of identities to vary at inference time. We then refine listener behavior using 12 hours of real two-person conversations. Our evaluation combines human and MLLM assessments of listener reactions with emotion metrics and a calibrated responsiveness metric that compares listener motion with real videos. Despite training on at most two people, AnyTalker generates three- and four-person conversations with listener interactions. Across the evaluated benchmarks, it matches or surpasses methods trained on more real multi-person data in lip synchronization, visual quality, and human judgments of naturalness.

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

Podcasts, interviews, live-streaming, and entertainment programs are fundamentally *conversational*: multiple people speak, listen, and react to one another. Audio-driven talking video generation has advanced rapidly on the single-person side (Lin et al., 2025b; Gan et al., 2025; Jiang et al., 2025b; Yang et al., 2025a; Cui et al., 2025; Chen et al., 2025a; Ji et al., 2025), powered by large video diffusion backbones (Blattmann et al., 2023; Yang et al., 2025b; Kong et al., 2024; Wan et al., 2025). Extending these models to multi-person conversations, however, has so far followed a data-hungry path: recent systems (Wang et al., 2025b; Huang et al., 2025; Kong et al., 2025) rely on hundreds to thousands of hours of curated multi-person data. Such data is expensive to collect and hard to annotate, since turn-taking, role-switching, and nonverbal cues like eye gaze all require fine-grained labels. Despite recent efforts to curate such data (Zhu et al., 2025; Chu & Patel, 2026), available multi-person corpora remain limited in scale, clip duration, and scenario diversity, with studio podcasts dominating their coverage. In contrast, the vast majority of publicly available audio-visual data (Zhang et al., 2021; Xie et al., 2022; Cui et al., 2025; Zhu et al., 2022; Chung et al., 2018) consists of open-domain single-person monologues.

To exploit this data asymmetry, we design AnyTalker around two complementary capabilities that can be learned from different data sources. The first is **multi-stream binding**: given several reference faces and audio tracks, drive each face with its own audio, and only that audio. The second is **interactivity**: the momentarily silent identities should behave like engaged listeners (gaze shifts, nods, eyebrow raises) rather than frozen portraits. We find that binding can be learned without real multi-person data. We horizontally concatenate two independent single-person clips while keeping their audio tracks separate. The resulting synthetic two-speaker samples teach the model per-identity audio binding and multi-speaker speaking patterns. We then use a small set of real two-person conversations to learn the cross-person reactions absent from these synthetic pairs. The resulting strategy uses roughly 2,000 hours of single-person data and only 12 hours of real two-person conversations. In terms of duration, this refinement set amounts to 12% of the 100 hours of real two-person conversations reported for the second training stage of MultiTalk (Kong et al., 2025). Our model generalizes beyond the two-person training regime at inference time.

To extend these capabilities beyond the two-person training setting, the architecture must support a variable number of identities at inference time. Prior binding mechanisms either hard-code speaker capacity or neglect the listener: MultiTalk (Kong et al., 2025) assigns discrete label embeddings whose range fixes the maximum number of speakers, Bind-Your-Avatar (Huang et al., 2025) routes at most two identities, and InterActHuman (Wang et al., 2025b), although architecturally designed for more than two speakers, produces listeners with little interactive behavior. We equip each DiT block with an *Audio-Face Cross Attention* (AFCA) module whose parameters are reused across identity-audio pairs. Each pair's output is masked to its face region before accumulation. AFCA supports this data strategy with shared parameters across identities and a common input format for synthetic and real conversations. Its parameter count is independent of identity count, with modest measured latency growth from two to four people.

Evaluating multi-person generation is itself an open problem. Existing benchmarks (Zhang et al., 2021; Xie et al., 2022) focus on single-person lip synchronization without assessing interactions between generated characters. The benchmark of Wang et al. (2025b) contains only one active speaker. To address this gap, we introduce *InteractivePerson* and a complementary three-axis protocol covering semantic reactions judged by humans and an MLLM, emotional plausibility, and ground-truth-calibrated motion responsiveness (described in Section 4).

Our contributions are:

- **A multi-person-data-efficient training strategy.** We show that concatenated single-person videos teach multi-stream audio binding and speaking patterns that transfer beyond the two-person training regime. Refinement on 12 hours of real two-person conversations improves listener behavior while limiting the need for real multi-person supervision.
- **A shared architecture for variable identity counts.** Each DiT block reuses its AFCA parameters across identity-audio pairs, allowing inference with more identities than seen in training. We evaluate two to four speakers quantitatively through lip synchronization, interactivity, and inference cost, with qualitative examples of larger groups.

- **Generation quality and listener behavior.** AnyTalker-14B achieves leading lip synchronization and visual quality on public benchmarks including HDTF, VFHQ, and EMTD. On InteractivePerson, human preferences and reaction annotations support improved listener naturalness, complemented by emotion metrics and calibrated motion responsiveness.

2 Related Work

Audio-driven talking video generation. Advances in diffusion models (Ho et al., 2020; Rombach et al., 2022; Ho & Salimans, 2022; Peebles & Xie, 2023) enabled end-to-end audio-driven portrait animation, pioneered by EMO (Tian et al., 2024) and followed by a rich line of work (Jiang et al., 2025a; Chen et al., 2025b; Xu et al., 2024; Lin et al., 2025a). These methods typically combine temporal attention (Guo et al., 2024), identity control via ReferenceNet (Hu, 2024), and audio conditioning through pre-trained speech encoders (Baevski et al., 2020; Radford et al., 2023) injected by cross-attention (Vaswani et al., 2017). Video foundation models (Blattmann et al., 2023; Yang et al., 2025b; Kong et al., 2024; Wan et al., 2025) pushed quality further, yielding systems with long-video generation (Yang et al., 2025a; Cui et al., 2025), full-body synthesis (Lin et al., 2025b), non-human subjects (Gao et al., 2025; Gan et al., 2025), and accurate lip synchronization (Ji et al., 2025). These models target single speakers. Applied naively to multi-person frames, they tend to synchronize all faces to the same audio.

Multi-person video generation. Dedicated multi-person systems bind audio to identities through embedding routers (Huang et al., 2025), region masks (Wang et al., 2025b), label rotary embeddings (Kong et al., 2025; Su et al., 2024), or face-aware spatial gating (Chen et al., 2025a). Playmate2 (Ma et al., 2026) uses mask-based classifier-free guidance at inference time. Several designs restrict identity count: MultiTalk fixes a maximum speaker capacity, and Bind-Your-Avatar is trained for two identities. Scalable alternatives such as InterActHuman and Playmate2 provide limited listener interactivity, while the released HunyuanVideo-Avatar implementation supports only single-identity inference. More importantly, training-based systems rely on hundreds to thousands of hours of curated multi-person data, while emerging datasets (Zhu et al., 2025; Chu & Patel, 2026) remain short and largely podcast-specific. AnyTalker instead composes binding supervision from open-domain single-person data and uses only 12 hours of real two-person conversations for interactivity refinement. SocialDirector (Ouyang et al., 2026), which is text-driven, and DreamID-Omni (Guo et al., 2026), which jointly generates audio and video without first-frame conditioning, fall outside our audio-driven, reference-image-conditioned setting.

3 Method

As illustrated in Fig. 2, AnyTalker combines a two-stage data strategy with a scalable audio-binding architecture. After formalizing the task in Section 3.1, we describe how single-person videos provide multi-speaker supervision and how real conversations refine interactivity in Section 3.2. We then introduce the architecture in Section 3.3, where each DiT block uses an *Audio-Face Cross Attention* (AFCA) module with parameters shared across identity-audio pairs to support this training strategy.

3.1 Task Formulation

The inputs are a reference image containing n identities, one audio track $a^{(k)}$ per identity, and a text prompt. The generated video should synchronize each identity k's lips with speech in $a^{(k)}$ and depict engaged listening during that identity's silent intervals. The number of identities is specified at inference time. We evaluate the same trained model with one to four identities.

3.2 The Data Strategy: Composing Multi-Person Supervision

Collecting real multi-person data for every group size is impractical, motivating us to separate the supervision for binding and interactivity. Multi-stream *binding* specifies which face responds to each audio stream and can be learned from independently recorded single-person clips. In contrast, *interactivity* concerns listeners' reactions to speakers and requires real conversations. These distinct requirements motivate the two training stages below.

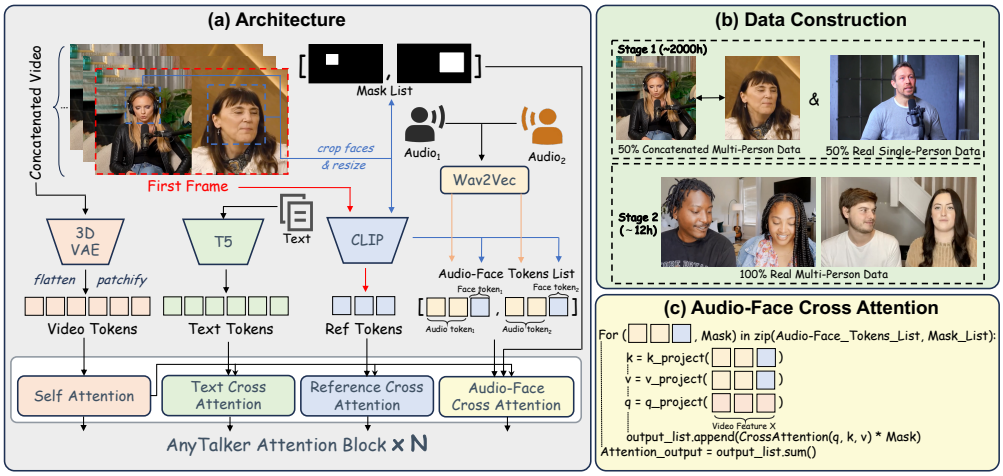


Figure 2: Overview of AnyTalker. (a) The generation backbone: a pre-trained I2V DiT (Wan et al., 2025) with an Audio-Face Cross Attention (AFCA) module in each block. (b) Stage 1 mixes ordinary single-person clips with *concatenated* pseudo-multi-person clips built from pairs of single-person videos. Stage 2 then refines interactivity on 12 hours of real two-person conversations. (c) Within each block, the same AFCA module is evaluated once per identity-audio pair. Its outputs are masked to the corresponding face regions and summed.

Stage 1: learning multi-speaker patterns from single-person data. We train on a 50/50 mixture of ordinary single-person clips and *concatenated* pseudo-multi-person clips. To construct the latter, we crop two independent videos around their face centers, concatenate them horizontally, and retain their separate audio tracks (Fig. 2b). We pair these samples with generic two-speaker text prompts. The resulting clips teach the model to bind each audio stream to the corresponding face and maintain plausible behavior when an identity is silent. Because the two clips are independent, their speech frequently overlaps, also teaching the model to handle simultaneously active speakers. We retain ordinary single-person clips to preserve lip accuracy and generation stability. The training-data ablation in Section 5.3 further evaluates their contribution. For batching, each concatenated sample joins a clip with the next-indexed clip in the batch, preserving the batch size used for ordinary clips. Data processing details are provided in Section B.

Stage 2: refine interactivity. Although stage 1 teaches audio binding, its independently recorded speakers cannot react to one another. To learn these missing interactions, we refine the model on real two-person conversations, which are more abundant and easier to curate than three- or four-person data. Restricting curation to two identities also enables reliable voice-to-face assignment and filtering using a 2×2 SyncNet score matrix (Li et al., 2024) (Fig. 6). We combine this check with detector-verified two-face presence (Deng et al., 2020), speaker diarization (Plaquet & Bredin, 2023), and optical-flow motion filtering (Karaev et al., 2024). The resulting subset contains 12 hours of real conversations. Real and concatenated samples share the same input format, so refinement requires no change to the model's conditioning interface. Despite using only pairs, this stage learns listening behaviors that transfer to larger groups at inference (Section 5.2).

3.3 Scalable Audio Binding

To use this supervision across different group sizes, AFCA processes each identity-audio pair with shared parameters within each DiT block. This design mirrors the structure of concatenated training clips, where each audio stream corresponds to a face in a separate region. The module computes an output for each pair, restricts it to the corresponding region, and sums the outputs.

Architecture overview. AnyTalker builds on the Wan I2V backbone and inherits its positional embeddings. The text, image, and audio encoders and the 3D VAE remain frozen. Video latents are patchified into a token grid with temporal length T', height H', and width W'. Flattening this

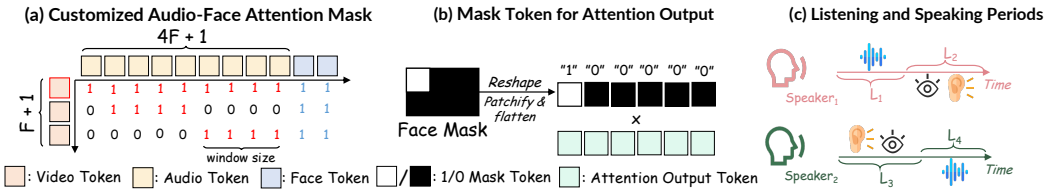


Figure 3: (a) Temporal alignment between video and audio features: each latent frame attends to four audio frame groups (the first attends globally). (b) The face mask is resized to the latent grid and flattened into a token-level mask applied to the AFCA output. (c) Speaking and listening phases used by the phase-level evaluation protocol.

grid gives $f_{\text{video}} \in \mathbb{R}^{L_{\text{video}} \times d}$, where $L_{\text{video}} = T' \times H' \times W'$ is the token count and d is the DiT hidden size. The VAE preserves the first frame and compresses subsequent frames by $4\times$, giving $T' = (T-1)/4 + 1$. For global conditioning, T5 (Raffel et al., 2020) encodes the text as f_{text}, while CLIP E_{CLIP} (Radford et al., 2021) produces the reference feature f_{ref}. For each identity k, Wav2Vec2 (Baevski et al., 2020) produces audio tokens $f^{(k)}_{\text{audio}}$, and E_{CLIP} encodes the first-frame face crop as $f^{(k)}_{\text{face}}$. A spatial mask $M^{(k)}_{\text{face}} \in [0,1]^{T' \times H' \times W'}$ marks the corresponding face region, while a fixed temporal mask M_{temporal} aligns video and audio tokens.

These per-identity inputs enter an AFCA module in each DiT block. Unlike label embeddings that fix speaker capacity (Kong et al., 2025) or routers trained for two identities (Huang et al., 2025), each module reuses its parameters across all identity-audio pairs. The resulting conditioning therefore accommodates a variable number of identities through the same sequence of encoding, masking, and accumulation operations.

Audio and face pair encoding. We separately project identity k's audio and face features into keys and values, then concatenate them along the token axis. The resulting pairs condition video queries:

$$\begin{aligned} K^{(k)} &= \text{Concat}\big(K^{(k)}_{\text{audio}}, K^{(k)}_{\text{face}}\big), \qquad V^{(k)} = \text{Concat}\big(V^{(k)}_{\text{audio}}, V^{(k)}_{\text{face}}\big), \\ A^{(k)} &= \text{MHCA}\big(Q_{\text{video}},\, K^{(k)},\, V^{(k)},\, M_{\text{temporal}}\big), \end{aligned} \tag{1}$$

where MHCA denotes multi-head cross attention and Q_{video} contains the projected video queries. The temporal mask aligns latent frames with local audio groups while keeping face tokens visible (Fig. 3a). Details are given in Section D.

Face-region masking. To localize this conditioning, we flatten the spatial mask and apply it element-wise to the attention output $A^{(k)}$:

$$M^{(k)}_{\text{token}} = \text{Flatten}(M^{(k)}_{\text{face}}), \qquad \text{AFCA}^{(k)} = M^{(k)}_{\text{token}} \odot A^{(k)}. \tag{2}$$

During training, the mask uses a time-independent box covering all detected positions of identity k (Fig. 3b). At inference, we expand the reference-frame box to construct the mask (Section C).

Iterative accumulation. We sum the masked outputs and apply an output projection. Omitting text and reference contributions, the update to DiT block i is

$$H'_i = H_i + \text{Proj}_i\Big(\sum\nolimits_{k=1}^{n} \text{AFCA}^{(k)}\Big), \tag{3}$$

where all n terms share the AFCA weights of block i, while different DiT blocks have distinct parameters. Thus, increasing n adds module evaluations without increasing the parameter count. The scalability evaluation in Section 5.2 measures the resulting inference cost and shows only modest latency growth from two to four identities. Implementation and parameter details are provided in Sections C and D.

4 EVALUATING MULTI-PERSON CONVERSATIONAL GENERATION

Standard talking-head evaluation measures the *speaker*'s lip synchronization and visual quality, leaving *listener* behavior less directly assessed. Our second training stage targets this behavior, motivating an evaluation of listener reactions and naturalness. Semantic judgments are costly, while motion statistics alone can mistake jitter for responsiveness. We therefore introduce a benchmark and a *multi-axis protocol* combining semantic assessments with motion scores calibrated against real videos.

4.1 THE MULTI-PERSON EVALUATION DATASET

We curate *InteractivePerson*, a set of real conversational clips involving two to four people, collected from YouTube. It comprises 100 clips: 60/20/20 contain 2/3/4 interacting people, respectively. Each clip is $\sim$10 seconds long, shot with a stable camera, and contains exactly 2 to 4 visible faces throughout. The clips contain rich listener reactions (mutual gaze, head turns, nods). Representative two-person examples are shown in Fig. 13, with both participants speaking during each clip. The first frame, used as the generation reference, always shows all participants facing the camera, so any subsequent interactive behavior must be *generated*, not copied from the input.

Annotation protocol. Speaking and listening intervals of each video are manually annotated. Five independent annotators segment each character's speaking/listening intervals. The maximum boundary disagreement is below 0.5 s, and final boundaries are the per-boundary mean. The phase-annotation example in Fig. 3c illustrates this segmentation for two people, with L_1, L_4 marking speaking phases and L_2, L_3 marking listening phases. All phase-level metrics below are duration-weighted and pooled over all characters, hence independent of the number of identities. For evaluation, all methods receive identical inputs: the reference frame, the separated audio tracks, and a text prompt.

4.2 A THREE-AXIS EVALUATION PROTOCOL

Axis 1: semantic reactions (human and MLLM judges). We define eight interactive reactions grounded in conversational behavior: *eyebrow raise, head turn, head nod, gaze shift, shoulder movement, gesture, stable mutual gaze (>1 s)*, and *listener mouth movement* (e.g., a responsive smile). We obtain two complementary assessments. (i) *Human raters*: 25 non-author AI researchers or product managers each evaluate roughly 40 randomly assigned comparison sets. Each set shows all applicable methods for one input in randomized, anonymous order. Raters mark each reaction's presence per video and select the most natural result, yielding detection rates and naturalness preference shares. (ii) *An MLLM judge*: Gemini-2.5-Pro detects the same eight reactions, following MLLM-based evaluation (Zheng et al., 2023) (prompt in Section E). These assessments measure visible listener behavior and perceived naturalness.

Axis 2: emotional plausibility. Using a seven-category emotion classifier (Serengil & Ozpinar, 2021), we report **Emo-Acc**, the label agreement between aligned generated and ground-truth face frames, and **Emo-Div**, the number of distinct labels per character. Emo-Acc measures classifier agreement rather than perceived emotional correctness. Emo-Div is reported alongside ground-truth diversity without a higher-is-better interpretation.

Axis 3: motion responsiveness. As an efficient automatic proxy, we track local eye-region keypoint motion during listening periods. We deliberately exclude shoulder and hand landmarks: their much larger, spatially unconstrained displacements make framewise motion distances unstable, whereas the human and MLLM judges in Axis 1 evaluate shoulder movement and gestures semantically. For the eye-region keypoint set $P = P_{\text{eye}}$ on the spatially normalized frames S_ℓ of listening phase ℓ, we define motion and its duration-weighted gap to ground truth as

$$\text{Motion}_\ell(P) = \frac{\sum_{j=1}^{|S_\ell|-1}\sum_{i=1}^{|P|}\|p_{i,j+1}-p_{i,j}\|_2}{(|S_\ell|-1)|P|}, \qquad \Delta\text{Resp} = \frac{\sum_\ell L_\ell \left|\text{Motion}_\ell^{\text{gen}} - \text{Motion}_\ell^{\text{GT}}\right|}{\sum_\ell L_\ell}. \quad (4)$$

Here ℓ ranges over all characters' listening phases, and L_ℓ is the phase duration. Lower ΔResp is better. Crucially, we do *not* treat larger motion as better. Prior work quantifies motion richness

Table 1: Multi-person results on InteractivePerson. Ground-truth Emo-Div is $4.52/4.67/4.59$ for 2/3/4 identities. Pref. Share denotes human naturalness preference. Latency is per denoising step for a 5-second 480P clip on one H200.

#IDs	Method	ΔResp ↓	Emo-Acc ↑	Emo-Div	Sync-C* ↑	FVD ↓	Pref. Share ↑	Latency (s)
2	Bind-Your-Avatar	0.51	44.82%	3.34	3.03	695.58	0.3%	N/A
	MultiTalk	0.27	61.79%	4.05	<u>6.88</u>	500.03	38.2%	7.51
	Playmate2	0.34	57.53%	3.91	5.10	545.22	12.1%	7.66
	AnyTalker-14B	**0.21**	**67.88%**	4.67	**7.02**	**424.15**	**49.4%**	**7.46**
3	Playmate2	0.38	49.65%	3.94	4.68	556.32	23.44%	9.18
	AnyTalker-14B	**0.23**	**62.56%**	4.72	**6.37**	**448.25**	**76.56%**	**7.55**
4	Playmate2	0.44	46.71%	3.95	4.40	579.03	21.76%	11.02
	AnyTalker-14B	**0.26**	**60.07%**	4.70	**6.14**	**469.84**	**78.24%**	**7.65**

through raw keypoint variance (Lin et al., 2025a; Tian et al., 2025), but such magnitude has no calibrated reference point. Real conversations provide a reference for motion magnitude, penalizing both insufficient and excessive movement relative to that reference. The eye-region landmark indices and spatial normalization are provided in Section E. Taking the per-phase difference *before* the duration-weighted average prevents overshoot in one phase from cancelling stillness in another, which a difference of pre-aggregated scores would allow. Head turns and eyebrow raises produce local landmark displacement (Fig. 8). This phase-level proxy does not establish reaction timing or semantic appropriateness, which requires complementary assessment.

5 Experiments

5.1 Experimental Setup

Training setup. Stage 1 uses about 2,000 hours of single-person data from public datasets (Zhang et al., 2021; Xie et al., 2022; Cui et al., 2025; Zhu et al., 2022; Chung et al., 2018; Chen et al., 2026; Zhang et al., 2026) and internet collection. Stage 2 uses the filtered 12-hour real-data subset (pipeline in Section B). We train two model sizes on Wan2.1-1.3B-Inp and Wan2.1-I2V-14B backbones. Text, audio, and image encoders and the 3D VAE are frozen. The DiT and AFCA layers are fully trainable. The learning rates for stages 1 and 2 are 2×10^{-5} and 5×10^{-6}, respectively, using AdamW (Loshchilov & Hutter, 2019) (hardware details in Section C).

Evaluation setup. Multi-person evaluation uses InteractivePerson with protocols described in Section 4, plus Sync-C*, i.e., Sync-C computed per speaking phase and duration-weighted with the same convention as Eq. (4): Sync-C$^* = \sum_s L_s$ Sync-C$_s / \sum_s L_s$, and FVD (Unterthiner et al., 2018). Emotion metrics cover all valid face frames. Face association and aggregation are detailed in Section E. Single-person evaluation uses HDTF and VFHQ with standard Sync-C, FID, FVD, and ID similarity, at the same number of single-person videos as MultiTalk. Scalability evaluation uses 60/20/20 clips for 2/3/4-person settings, respectively.

For multi-person, we compare against MultiTalk (Kong et al., 2025), Bind-Your-Avatar (Huang et al., 2025), and Playmate2 (Ma et al., 2026), whose mask-based classifier-free guidance also scales to three or more characters. We exclude InterActHuman (Wang et al., 2025b) because its official release does not provide multi-person inference. For single-person generation, we compare against EchoMimic (Chen et al., 2025b), Hallo3 (Cui et al., 2025), Sonic (Ji et al., 2025), FantasyTalking (Wang et al., 2025a), StableAvatar (Tu et al., 2025), and OmniHuman-1.5 (Jiang et al., 2025b), as well as MultiTalk and Playmate2.

5.2 Results

Multi-person comparison. Table 1 compares lip synchronization, visual quality, listener behavior, and inference latency on InteractivePerson. For Bind-Your-Avatar's 49-frame limit, we condition each subsequent segment on the preceding segment's final frame until its duration matches the other methods (Section C). Its per-step latency remains incomparable across these segments. Across the

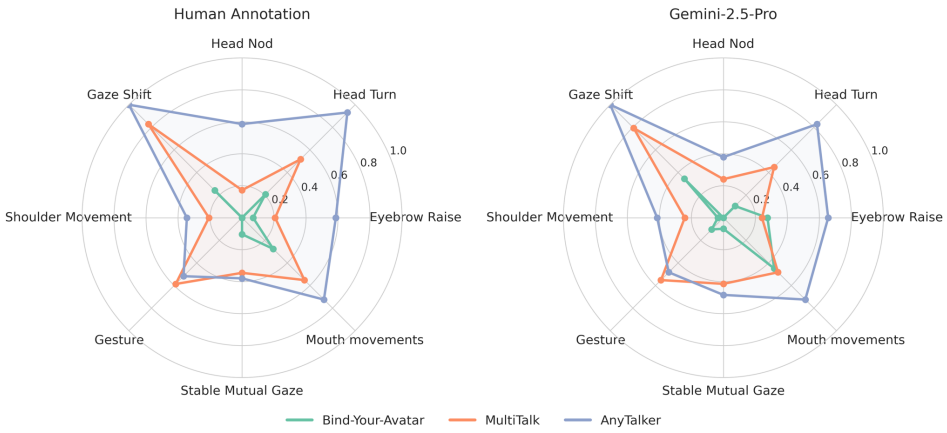


Figure 4: Detection rates of the eight semantic listener reactions across methods, judged by human annotators (left) and Gemini 2.5 Pro (right).



Figure 5: Qualitative results of AnyTalker in two- to four-person settings.

evaluated settings, AnyTalker has the lowest ΔResp and FVD, the highest Emo-Acc and Sync-C*, and the largest naturalness preference share. Its emotion diversity is also closest to ground truth. The reaction radar (Fig. 4) shows higher detection rates for gaze shifts, head nods, head turns, and eyebrow raises, the listener behaviors targeted by refinement.

Scaling the number of identities. Training uses at most two identities. Table 1 tests generalization to three and four. Lip synchronization drops mildly with added identities, responsiveness stays close to ground truth, and emotion diversity tracks it. MultiTalk and Bind-Your-Avatar support only two identities. Playmate2 uses a separate guidance pass per identity. AnyTalker instead repeats AFCA within each block, with measured latency increasing from 7.46 to 7.65 s between two and four identities. Fig. 5 shows two- to four-person results with InteractivePerson and AIGC-generated references (Team, 2024). Larger groups, non-human characters, and comparisons appear in Fig. 11 and Section G and our demo page[1].

Single-person generation. A multi-person strategy should not tax single-person quality. On HDTF and VFHQ (Table 2), AnyTalker-14B achieves best or tied-best scores on every metric against eight baseline systems, and the 1.3B model outperforms EchoMimic and StableAvatar on lip synchronization by large margins. AnyTalker-14B also leads on three of four metrics on the half-body EMTD benchmark (Meng et al., 2025) (Section F).

5.3 Ablation Studies

Data strategy. We compare training-data combinations using the 1.3B model on two-person InteractivePerson in Table 3. Adding concatenated clips to ordinary single-person training improves

[1] https://completely-anonymous-z.github.io/audio-driven-demo/

Table 2: Single-person results on HDTF and VFHQ. OmniHuman-1.5* is the "Master Mode" version accessed via the JiMeng platform (ByteDance, 2025).

Method	Backbone	HDTF				VFHQ			
		Sync-C ↑	FID ↓	FVD ↓	ID ↑	Sync-C ↑	FID ↓	FVD ↓	ID ↑
FantasyTalking	Wan 2.1	3.97	14.93	166.79	0.93	3.57	24.83	272.13	0.94
StableAvatar	Wan 2.1	4.11	14.67	166.44	0.91	3.53	22.91	275.73	0.88
EchoMimic	Stable Diffusion v1.5	5.23	61.53	381.55	**0.95**	4.87	58.72	486.75	0.86
Hallo3	CogVideoX	7.53	17.12	195.61	0.91	6.32	41.26	371.24	0.91
Sonic	Stable Video Diffusion	7.81	52.96	286.12	**0.95**	7.71	36.68	385.37	0.89
OmniHuman-1.5*	Seedance 1.0	7.23	35.26	173.23	0.90	7.67	35.36	283.39	0.90
MultiTalk	Wan 2.1	8.91	13.54	162.58	0.93	7.77	24.25	243.66	0.94
Playmate2	Wan 2.1	8.64	13.04	147.66	0.92	7.41	25.01	259.79	0.92
AnyTalker-1.3B	Wan 2.1	6.92	14.22	211.32	0.92	5.99	21.41	256.43	0.92
AnyTalker-14B	Wan 2.1	**9.27**	**12.64**	**142.56**	**0.95**	**7.92**	**18.65**	**241.52**	**0.95**

Table 3: Data-strategy ablation on two-person InteractivePerson. RS, CM, and RM denote real single-person, concatenated multi-person, and real multi-person data. Superscripts indicate the training stage.

Data	ΔResp ↓	Emo-Acc ↑	Sync-C* ↑
CM^1	0.49	39.94%	3.34
RS^1	0.53	37.72%	4.39
RS^1+CM^1	0.47	42.25%	4.97
RS^1+RM^2	0.35	47.18%	3.88
RS^1+CM^1+RM^2	**0.32**	**52.83%**	**5.02**

Table 4: Design ablations on HDTF (stage-1, 1.3B). Baseline uses basic audio attention. w/o AFCA retains CM and face masks but omits face-token concatenation. CM denotes concatenated training data.

Setting	Sync-C ↑	FVD ↓	ID ↑
Baseline	5.56	167.24	0.91
w/o AFCA	6.84	184.36	0.89
w/o Face Mask	5.98	175.82	0.89
w/o CM	6.36	187.52	0.91
Full (stage-1)	**7.05**	**160.61**	**0.92**

every metric, whereas training on concatenated clips alone reduces lip accuracy and generation stability. Refining the mixed-data model on the 12-hour real-data subset reduces the responsiveness gap from 0.47 to 0.32 and raises Sync-C* from 4.97 to 5.02. Refinement without concatenated pretraining performs worse on every metric than the full two-stage strategy. Visual comparisons before and after refinement in Section H further illustrate the improvement in eye contact.

Module and data ablations. Table 4 evaluates the stage-1 1.3B model on HDTF. The w/o AFCA variant uses ordinary audio-video cross attention with face masks and concatenated training data, omitting only face-token concatenation. The other variants remove face-region masking or concatenated training data (w/o CM). The full model achieves the best Sync-C, FVD, and ID scores among these settings.

6 Conclusion

We showed that concatenated single-person clips provide supervision for multi-stream audio binding, while refinement on 12 hours of real two-person conversations improves listener behavior. AFCA reuses parameters across identity-audio pairs within each DiT block, supporting variable identity counts. A model trained on at most two people generates three- and four-person conversations, with modest measured latency growth and no additional parameters. Our InteractivePerson benchmark and evaluation protocol combine human and MLLM assessments of listener reactions with emotion statistics and motion measurements calibrated against real conversations. Across the evaluated benchmarks, AnyTalker matches or surpasses the evaluated methods in synchronization and visual quality, while human assessment favors the naturalness of its conversational videos. This strategy of composing multi-entity supervision from single-entity data and refining interactions with a small real corpus may extend beyond talking heads.

AI USE STATEMENT

Large Language Models (LLMs) were used to aid in the code writing and manuscript polishing. Specifically, the usage includes refining the language, improving readability, and ensuring clarity in the paper. It is important to note that LLMs were not involved in the ideation, research methodology, or experimental design.

REPRODUCIBILITY STATEMENT

We will release the inference code as open source as soon as possible. For now, please refer to our demo page[2] for generation results. Experimental settings, implementation details, and hyperparameter configurations are documented in Sections C and E.

ETHICS STATEMENT

Our training corpus combines existing public research datasets (Zhang et al., 2021; Xie et al., 2022; Cui et al., 2025; Zhu et al., 2022; Chung et al., 2018) with videos collected from public internet platforms. Internet collection follows a selection protocol restricted to publicly posted content. For the InteractivePerson evaluation benchmark, we prioritize data of public figures in public-facing settings. Distribution of evaluation data will be contingent on explicit consent from the original creators. Data whose creators decline will be replaced with self-recorded material or anonymized.

REFERENCES

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

ByteDance. Jimeng platform. `https://jimeng.jianying.com/ai-tool/home?type=digitalHuman`, 2025.

Shunian Chen, Hejin Huang, Yexin Liu, Zihan Ye, Pengcheng Chen, Chenghao Zhu, Michael Guan, Rongsheng Wang, Junying Chen, Guanbin Li, Ser-Nam Lim, Harry Yang, and Benyou Wang. Talkvid: A large-scale diversified dataset for audio-driven talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 3492–3500, 2026.

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*, 2025a.

Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2403–2410, 2025b.

Ernie Chu and Vishal M Patel. Face-to-face: A video dataset for multi-person interaction modeling. *arXiv preprint arXiv:2603.14794*, 2026.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, pp. 1086–1090, 2018.

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.

[2] `https://completely-anonymous-z.github.io/audio-driven-demo/`

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*, pp. 4690–4699, 2019.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5203–5212, 2020.

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.

Xu Guo, Fulong Ye, Qichao Sun, Liyang Chen, Bingchuan Li, Pengze Zhang, Jiawei Liu, Songtao Zhao, Qian He, and Xiangwang Hou. Dreamid-omni: Unified framework for controllable human-centric audio-video generation. *arXiv preprint arXiv:2602.12160*, 2026.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8163, 2024.

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.

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, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. In *The Thirteenth International Conference on Learning Representations*, 2025a.

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*, 2025b.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European Conference on Computer Vision*, pp. 18–35. Springer, 2024.

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. In *Advances in Neural Information Processing Systems*, volume 38, pp. 70990–71013, 2025.

Chunyu Li, Chao Zhang, Weikai Xu, Jingyu Lin, Jinghui Xie, Weiguo Feng, Bingyue Peng, Cunjian Chen, and Weiwei Xing. Latentsync: Taming audio-conditioned latent diffusion models for lip sync with syncnet supervision. *arXiv preprint arXiv:2412.09262*, 2024.

Gaojie Lin, Jianwen Jiang, Chao Liang, Tianyun Zhong, Jiaqi Yang, Zerong Zheng, and Yanbo Zheng. Cyberhost: A one-stage diffusion framework for audio-driven talking body generation. In *The Thirteenth International Conference on Learning Representations*, 2025a.

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*, pp. 13848–13858, 2025b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Xingpei Ma, Shenneng Huang, Jiaran Cai, Yuansheng Guan, Shen Zheng, Hanfeng Zhao, Qiang Zhang, and Shunsi Zhang. Training-free multi-character audio-driven animation via diffusion transformer with reward feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7818–7826, 2026.

Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5489–5498, 2025.

Liangyang Ouyang, Ruicong Liu, Caixin Kang, Yifei Huang, and Yoichi Sato. Socialdirector: Training-free social interaction control for multi-person video generation. *arXiv preprint arXiv:2605.10079*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4172–4182, 2023.

Alexis Plaquet and Hervé Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. In *INTERSPEECH*, pp. 3222–3226, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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*, pp. 28492–28518. PMLR, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *International Conference on Engineering and Emerging Technologies (ICEET)*, 2021.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Naoya Takahashi and Yuki Mitsufuji. Multi-scale multi-band densenets for audio source separation. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 21–25. IEEE, 2017.

Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. Technical report, 2024. URL https://github.com/Kwai-Kolors/Kolors/blob/master/imgs/Kolors_paper.pdf.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pp. 244–260. Springer, 2024.

Linrui Tian, Siqi Hu, Qi Wang, Bang Zhang, and Liefeng Bo. Emo2: End-effector guided audio-driven avatar video generation. *arXiv preprint arXiv:2501.10687*, 2025.

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, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *CoRR*, abs/1812.01717, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

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.

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*, pp. 9891–9900, 2025a.

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*, 2025b.

Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 657–666, 2022.

Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, et al. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. *arXiv preprint arXiv:2508.14033*, 2025a.

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 *The Thirteenth International Conference on Learning Representations*, 2025b.

Youliang Zhang, Zhaoyang Li, Duomin Wang, Jiahe Zhang, Deyu Zhou, Zixin Yin, Xili Dai, Gang Yu, and Xiu Li. Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. In *International Conference on Learning Representations*, 2026.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2023.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European Conference on Computer Vision*, pp. 650–667. Springer, 2022.

Zeyu Zhu, Weijia Wu, and Mike Zheng Shou. Multi-human interactive talking dataset. *arXiv preprint arXiv:2508.03050*, 2025.

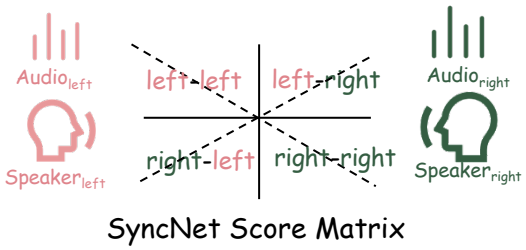


Figure 6: SyncNet score matrix used for voice-to-face assignment in two-person data.

A Appendix

B Data Processing Details

B.1 Auxiliary Models

To segment sentences and chunk audio tracks, we use speaker-diarization-3.1 (Plaquet & Bredin, 2023)[3] with Whisper (Radford et al., 2023). Vocal separation and features use Kim_Vocal_2 (Takahashi & Mitsufuji, 2017) and Wav2Vec2 (Baevski et al., 2020). Face detection and identity features use RetinaFace (Deng et al., 2020) and ArcFace (Deng et al., 2019) via InsightFace. Audio-visual synchronization is scored by the pre-trained SyncNet of LatentSync (Li et al., 2024). Text prompts use fixed generic two-speaker templates during stage-1 training and fixed identity-count-matched templates at inference. All prompts are encoded with T5 (Raffel et al., 2020). Camera-motion filtering uses CoTracker (Karaev et al., 2024).

B.2 Single-Person Data

A valid single-person clip is a 24-fps video continuously exhibiting one identifiable face, with speech audio synchronized to lip movements. We eliminate recordings with pronounced camera shake or scene changes via CoTracker, verify single-face presence with RetinaFace, obtain sentence-level segments via diarization, and subdivide utterances longer than 5 s into semantically coherent units with Whisper, discarding fragments with overlapping speakers. Remaining segments average $\approx$2 s. Consecutive segments are concatenated so clip lengths follow a Gaussian with mean 4 s (std 0.5 s). Vocal tracks are separated and encoded, SyncNet-scored clips below threshold are removed, and retained clips are transcribed and annotated with frame-level face boxes. This yields approximately 2,000 h of 480P single-person data.

B.3 Two-Person Data

The two-person pipeline follows the single-person one, including CoTracker camera-motion filtering, with three additional checks. (1) **Face count**: most frames must contain exactly two faces. (2) **Speaker activity**: diarization is restricted to $\{$both active, one active$\}$. (3) **Spatial consistency**: left-to-right face ordering must remain unchanged, with identity swaps detected and rejected via InsightFace. Voice-to-face correspondence is established through a 2×2 SyncNet confidence matrix (Fig. 6). Clips whose two largest scores are off-diagonal are discarded, and the single-person synchrony threshold is applied. Filtering yields a clean 12-hour subset of real two-person conversations for interactivity refinement.

B.4 Concatenation Strategy

Naively resizing high-resolution sources and stacking them yields tiny faces. We instead locate the face center of each candidate clip and expand a minimal crop of $(480, 416)$ around it (480P frames are $(480, 832)$), then randomly enlarge the window while preserving the aspect ratio and staying

[3] https://huggingface.co/pyannote/speaker-diarization-3.1

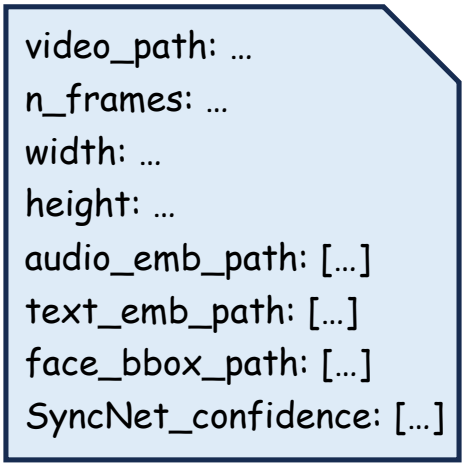


Figure 7: Properties of training data.

inside the frame. The resulting crops keep faces large, enabling accurate lip-to-audio mapping. Each processed sample stores paths to decoded video, cleaned audio, and pre-extracted features in a dictionary item with an identical key structure for single- and two-person data (Fig. 7).

B.5 Training and Evaluation Separation

We use ArcFace identity embeddings (Deng et al., 2019) to screen for identity overlap between the training data and evaluation sets. Training identities with cosine similarity of at least 0.3 to any test identity are excluded. After filtering, the maximum cosine similarity between each test identity and the retained training identities is strictly below 0.3. The evaluation data provide clear facial views, and training clips without detectable faces are removed during preprocessing. ArcFace embeddings are therefore available for all identities in this screening.

C Implementation Details

Training. The 14B model trains on 32 H200 GPUs: stage 1 uses global batch size 32 for 75K optimizer steps, followed by stage 2 with batch size 16 for 50K further steps. The 1.3B model trains on 8 H200 GPUs: stage 1 uses a global batch size 48 for 50K optimizer steps, followed by 50K stage-2 steps. Stage 2 uses a warm-up schedule. Encoders and VAE are frozen throughout. The DiT and AFCA layers are fully trainable.

Parameter accounting. The name AnyTalker-14B follows its Wan I2V backbone family. We count scalar parameters in the conditioned DiT checkpoint, including biases and affine normalization parameters. The 40 AFCA attention modules use hidden width $d = 5{,}120$ and audio width $d_a = 768$. Their query, audio key/value, face key/value, and output projections contribute $40(4d^2+2d_ad+6d) \approx 4.51$B parameters. Table 5 lists the input normalization and audio/face feature projectors separately, giving a total of 20.99B parameters for the conditioned DiT. Frozen T5, CLIP, Wav2Vec2, and VAE weights are excluded. Each block has its own AFCA weights, which are reused across identity-audio pairs, so the parameter count does not depend on the number of identities.

Inference. All baselines use their released implementations and default hyperparameters. To match evaluation durations, we extend Bind-Your-Avatar beyond its 49-frame generation limit by using each generated segment's final frame as the reference image for the next segment. Generation continues with the subsequent audio portion until the output matches the duration used for the other methods. Quality and human assessments use these extended outputs. Its per-step latency is marked N/A because generation requires multiple conditioned segments rather than one pass over the target duration. Text-conditioned baselines receive "`this person is talking`" on single-person benchmarks. For multi-person inputs, all methods receive "`The two persons are talking`

Table 5: Parameter breakdown for AnyTalker-14B. Counts include biases and affine normalization parameters. Frozen auxiliary models are excluded.

Conditioned DiT component	Parameters
AFCA attention projections (40 blocks)	4,510,105,600
AFCA input normalization (40 blocks)	409,600
Audio feature projector	74,214,400
Face feature projector	8,211,200
Remaining DiT parameters	16,394,878,784
Total conditioned DiT	20,987,819,584

`to each other.`", with `two` replaced by `three` or `four` to match the number of identities. AnyTalker uses classifier-free guidance (Ho & Salimans, 2022) at scale 4.0. Face masks are constructed solely from the reference frame. Given n input audio streams, InsightFace retains the n largest face detections and orders them from left to right to establish the audio-to-identity correspondence. Each detected rectangle is converted to an image-clipped square binary mask, expanded about its center by factors of 1.2 in width and 1.1 in height, bilinearly resized to the latent token grid, and repeated unchanged over all latent frames. These static masks gate the corresponding AFCA outputs without tracking generated faces. Partial overlaps between expanded masks are permitted, as in training.

D AFCA Implementation Details

This section expands the AFCA formulation in Section 3.3. We describe the conditioning features, temporal alignment, and spatially masked accumulation using a common notation across model sizes.

D.1 Video and Conditioning Features

Let B be the batch size and let the output contain $T = 4m+1$ frames at resolution $H \times W$. The VAE produces 16-channel latents with $F = m+1$ temporal positions and spatial size $(H/8)\times(W/8)$. The DiT concatenates the noisy latents, reference-video latents, and four conditioning-mask channels to form a 36-channel input. A $1 \times 2 \times 2$ patch embedding maps this input to

$$X \in \mathbb{R}^{B\times L\times d}, \qquad L = FH'W', \quad H' = H/16, \quad W' = W/16. \tag{5}$$

Each identity mask has shape $B \times 1 \times F \times H' \times W'$ before flattening and $B \times L \times 1$ afterward. The final singleton dimension broadcasts across the d feature channels.

Table 6: Backbone dimensions used by AFCA. Each block reuses its conditioning parameters across identities.

Configuration	1.3B	14B
DiT blocks	30	40
Hidden width d	1,536	5,120
Attention heads	12	40
Channels per head	128	128
Projected face tokens	$257 \times 1{,}536$	$257 \times 5{,}120$

Audio features. Each audio stream is resampled to 16 kHz. Wav2Vec2 convolutional features are linearly interpolated to the video frame rate before entering its transformer. We retain the 12 transformer-layer outputs, each with 768 channels. For every video frame, we stack the features at offsets $\{-2, -1, 0, 1, 2\}$, clipping indices at the sequence boundaries. The resulting tensor for one identity has shape $B \times T \times 5 \times 12 \times 768$.

Face features. CLIP encodes each reference-frame face crop at 224×224 resolution into 257 tokens with 1,280 channels. A learned projection maps these tokens to width d using LayerNorm,

a $1{,}280 \to 1{,}280$ linear layer, GELU, a $1{,}280 \to d$ linear layer, and a final LayerNorm. The face projector is shared across identities and is separate from the projector for the full reference image.

D.2 Audio Projection and Temporal Alignment

Audio projection. The learned audio projector converts contextual Wav2Vec2 features into attention tokens. It uses two hidden layers of width 512 with ReLU activations, followed by an output layer producing 32 tokens per temporal group and LayerNorm on the output tokens. The audio and face features are then mapped to keys and values of width d for joint attention. We denote their token counts by L_a and L_f, with $L_f = 257$ for the CLIP face representation.

Temporal alignment. Following the frame-group notation in Fig. 3a, latent frame $u > 0$ is associated with the audio groups for video frames $4u - 3$ through $4u$, using zero-based indices. The first latent frame has global audio visibility in this formulation. Face tokens remain visible to every video query. The Boolean attention mask is broadcast across heads and has shape

$$B \times 1 \times L \times (L_a + L_f). \tag{6}$$

The audio columns encode temporal visibility, while the face columns are always enabled. Spatial queries within the same latent frame share the same temporal visibility pattern. Each frame group can contain multiple projected audio tokens, so the four groups in the diagram represent temporal alignment units rather than four individual attention tokens.

D.3 Attention, Spatial Masking, and Residual Updates

Joint audio and face attention. Audio and face features use different learned key and value projections. These projections are reused for all identities within a block. Writing $C_a^{(k)}$ and $C_f^{(k)}$ for identity k's projected audio and face features, the key and value sequences are

$$\begin{aligned} K^{(k)} &= \mathrm{Concat}\Big(C_f^{(k)} W_K^f,\ C_a^{(k)} W_K^a\Big), \\ V^{(k)} &= \mathrm{Concat}\Big(C_f^{(k)} W_V^f,\ C_a^{(k)} W_V^a\Big). \end{aligned} \tag{7}$$

The implementation places face tokens before audio tokens. This ordering is equivalent to Eq. (1) when the attention-mask columns follow the same order. Each projection outputs width d, which is split into heads of width 128. Attention uses the usual $1/\sqrt{128}$ dot-product scaling. AFCA concatenates the audio and face tokens so that both contribute to the same attention softmax.

Ablation definitions. In Table 4, w/o AFCA removes face-token concatenation and uses ordinary cross attention between video queries and audio keys and values. It retains the same mixture of single-person and concatenated multi-person training clips as the full model. Each identity's attention output is still gated by its face mask before accumulation, so this variant preserves spatial audio binding. The w/o CM variant instead excludes concatenated multi-person clips from training while retaining AFCA and face-region masks. The w/o Face Mask variant removes spatial gating of the attention outputs while retaining AFCA and concatenated training data.

Normalization and residual update. Video features are normalized before cross attention. The block normalization uses $\epsilon = 10^{-6}$. Each identity's attention output is spatially masked, and the outputs are summed before the $d \to d$ output projection and residual addition in Eq. (3). This conditioning operates within each DiT block alongside text and reference-image conditioning. Attention and output dropout are zero. The model constructor uses Xavier uniform initialization for linear projections with zero biases. The AFCA output projection is not zero-initialized, and there is no additional learned gate on this branch.

Mask construction and accumulation. At inference, the detector selects the n largest faces and orders them from left to right to match the audio list. Each face box is converted to an image-clipped square. Its binary mask is expanded by factors of 1.2 horizontally and 1.1 vertically, then bilinearly resized to $H' \times W'$ with `align_corners=False`. The resized mask is repeated over all F latent frames and flattened to gate the corresponding attention output. Bilinear resizing can produce

fractional mask values at boundaries. During training, the mask instead uses a static box covering the identity's detected positions throughout the clip.

Masking occurs before the shared output projection. Outputs are summed without division by the number of identities or by the mask sum. Non-overlapping regions receive only their assigned identity's masked output. Where masks overlap, their contributions add. The identity loop evaluates every pair against the same video queries, so it does not update the video state between pairs. Different DiT blocks have distinct parameters. Increasing identity count adds attention evaluations, while the number of learned parameters remains fixed. The backbone retains global video self-attention, allowing information to propagate across face regions between blocks.

E Evaluation Protocol Details

MLLM judge. The judge receives each clip together with the following instruction, which defines the eight reactions and requires a binary decision with a short, visually grounded justification for each one.

```
You are a strict video annotator for multi-person conversations.
Determine whether each semantic listener reaction below visibly occurs
at least once in the entire video.

A listener is a visible person who is not speaking at that moment.
Infer speaking status jointly from audible speech and visible lip motion.
Count only actions performed while the person is a listener. An action
must be a visible change over time; do not infer it from a static pose.

Reaction definitions:
1. eyebrow_raise: one or both eyebrows visibly move upward.
2. head_turn: the listener changes head orientation toward or away from
   another participant. Do not count camera motion.
3. head_nod: a clear vertical dip-and-return of the head that functions
   as a conversational response. Do not count camera or body shake.
4. gaze_shift: the listener visibly redirects gaze between another
   participant, the camera, or another target.
5. shoulder_movement: a clear shrug, lift, roll, or responsive shoulder
   displacement. Do not count breathing or global camera motion.
6. gesture: an intentional communicative hand or arm movement, such as
   pointing, waving, or an illustrative response. Ignore incidental
   fidgeting.
7. stable_mutual_gaze: two participants look at each other continuously
   for more than one second, with at least one acting as a listener.
8. listener_mouth_movement: a visible non-speaking mouth response from
   a listener, such as a smile, laugh, or brief backchannel movement.
   Do not count ordinary speaking lip motion.

Rules:
- Inspect the full clip. The eight labels are not mutually exclusive.
- Count only clearly visible events; if evidence is ambiguous, use false.
- Do not use scene cuts, camera motion, or movement by another person as
  evidence for the target listener.
- For each evidence field, write exactly one short sentence naming the
  visible person and the observed change. If false, write
  "No clear instance is visible."
- Return only valid JSON with exactly the schema and key order below.
  Do not add Markdown, analysis, scores, or additional keys.

{
  "eyebrow_raise": {"present": false, "evidence": "..."},
  "head_turn": {"present": false, "evidence": "..."},
  "head_nod": {"present": false, "evidence": "..."},
  "gaze_shift": {"present": false, "evidence": "..."},
  "shoulder_movement": {"present": false, "evidence": "..."},
  "gesture": {"present": false, "evidence": "..."},
```

```
  "stable_mutual_gaze": {"present": false, "evidence": "..."},
  "listener_mouth_movement": {"present": false, "evidence": "..."}
}
```

Videos are supplied with audio and uniformly sampled at 5 fps, with the instruction preceding the video input. We use `gemini-2.5-pro` with temperature 0, top-$p = 1$, top-$k = 1$, seed 42, one candidate, and a maximum of 2,048 output tokens. Malformed or schema-inconsistent responses are retried once. One accepted response is retained per clip, without self-consistency voting. The evidence strings are retained for auditing but are not scored. For reaction r, the MLLM detection rate is the mean of its `present` indicator over all evaluated clips.

Human study. Twenty-five non-author participants, all AI researchers or product managers, each evaluate roughly 40 randomly assigned comparison sets. Each set corresponds to one input and identity-count setting and concatenates the outputs of all applicable methods in randomized order, without exposing method names. For every constituent video, participants mark the presence or absence of the same eight reactions defined for the MLLM judge and then select the single video they consider most natural. The human detection rate for each reaction is the fraction of positive binary annotations. Naturalness preference share is the fraction of best-of-set selections received by each method within an identity-count setting, using the completed comparisons for that setting as the denominator. This is a choice among all methods in a set, rather than a collection of pairwise comparisons.

Emotion metrics. We use the DeepFace emotion classifier (Serengil & Ozpinar, 2021) with its seven-label space {*angry, disgust, fear, happy, sad, surprise, neutral*}. Generated and ground-truth videos are paired by frame index and evaluated up to the length of the shorter sequence. Within each frame, face crops are associated by the fixed identity ordering of the benchmark. A frame-identity pair is retained only when the corresponding face is detected in both videos. For each retained crop, DeepFace produces a seven-way probability vector, whose argmax is used as the frame-level emotion label. Writing $\hat{y}^{\mathrm{gen}}_{vkt}$ and $\hat{y}^{\mathrm{GT}}_{vkt}$ for the generated and ground-truth labels of identity k in video v at frame t, we compute

$$\text{Emo-Acc} = \frac{1}{|\Omega|} \sum_{(v,k,t)\in\Omega} \mathbf{1}\left[\hat{y}^{\mathrm{gen}}_{vkt} = \hat{y}^{\mathrm{GT}}_{vkt}\right], \tag{8}$$

where Ω contains all valid aligned frame-identity pairs. The metric measures agreement between classifier predictions over both speaking and listening frames. It does not directly measure human judgments of emotional appropriateness. For emotion diversity, we count the distinct frame-level labels displayed by each identity over the entire video,

$$D_{vk} = \left|\{\hat{y}^{\mathrm{gen}}_{vkt} : t \in \Omega_{vk}\}\right|, \tag{9}$$

and report **Emo-Div** as the mean of D_{vk} over videos and identities. The ground-truth diversity is computed identically and reported as a reference point. Emo-Div is therefore not treated as a higher-is-better metric.

Calibrated responsiveness. Each detected face is mapped to a common 224×224 crop coordinate system, and the per-phase keypoint motion (Eq. (4)) is computed only over the local eye region using Euclidean (ℓ_2) displacement. Facial landmarks are extracted on each frame with the InsightFace 106-point detector. Using zero-based indices in this layout, the eye-region set contains both eyebrows and both eyes: $P_{\mathrm{eye}} = \{43, 48, 49, 50, 51, 101, 102, 103, 104, 105\} \cup \{35, 36, 37, 39, 41, 42, 89, 90, 91, 93, 95, 96\}$. We do not include shoulder or hand landmarks in this displacement metric: their large-amplitude motion is substantially more sensitive to pose and localization errors, so shoulder movement and gestures are instead evaluated semantically by the human and MLLM judges. The score measures the absolute difference in mean displacement within matched listening phases. Neither larger nor smaller motion is intrinsically better. Different motion trajectories can have the same phase-level mean, so this proxy alone cannot establish reaction timing, distinguish all jitter from meaningful motion, or demonstrate dependence on another speaker's audio. Fig. 8 visualizes the tracked landmarks: head turns toward the speaker or eyebrow raises produce measurable local displacement, whereas sustained stillness produces little.

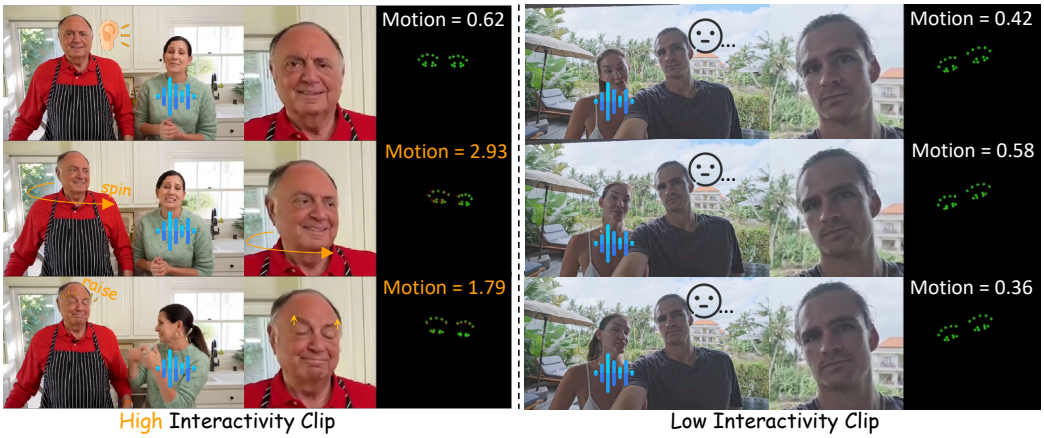


Figure 8: InteractivePerson examples with raw eye-region motion (px). Original videos are shown on the left, with face crops and tracked landmarks on the right. Head turns or eyebrow raises produce larger local displacement than stillness. $\Delta\mathrm{Resp}$ measures deviation from ground-truth phase-level motion (lower is better).

Table 7: Quantitative results on the half-body EMTD benchmark (Meng et al., 2025).

Method	**Sync-C** ↑	**FID** ↓	**FVD** ↓	**ID** ↑
FantasyTalking (Wang et al., 2025a)	3.76	67.66	818.71	0.76
EchoMimic v2 (Meng et al., 2025)	6.27	63.43	671.18	0.76
MultiTalk (Kong et al., 2025)	8.38	64.71	787.99	**0.79**
AnyTalker-1.3B	5.83	56.01	789.59	0.74
AnyTalker-14B	**8.45**	**50.61**	**664.58**	0.77

F Single-Person Benchmarks

The HDTF/VFHQ comparison is in Table 2 (main text). On the half-body EMTD benchmark (Table 7), AnyTalker-14B leads on three of four metrics. Qualitative single-person comparisons are shown in Fig. 9.

G Additional Multi-Person Results

Fig. 11 directly compares AnyTalker with prior systems under matched reference images, prompts, and per-identity audio streams. In the two-person examples, the listener generated by MultiTalk remains largely inert and Bind-Your-Avatar produces comparatively static expressions, whereas AnyTalker exhibits clearer gaze and head interactions. Because MultiTalk and Bind-Your-Avatar do not support three or more identities, the lower block compares AnyTalker with Playmate2. Across both three- and four-person examples, AnyTalker produces more visible reactions from non-speaking participants.

Fig. 12 shows additional multi-person generations spanning real photographs, AIGC images, and cartoons, with varied identity counts. Fig. 13 shows representative InteractivePerson clips.

H Effect of Interactivity Refinement

As shown in Fig. 14, stage-2 refinement lets identities exchange substantially more natural eye contact. A model trained solely on stage-1 data drives the assigned faces but shows limited listener reactions in these examples. Refinement on 12 hours of real two-person conversations produces more visible gaze interaction.

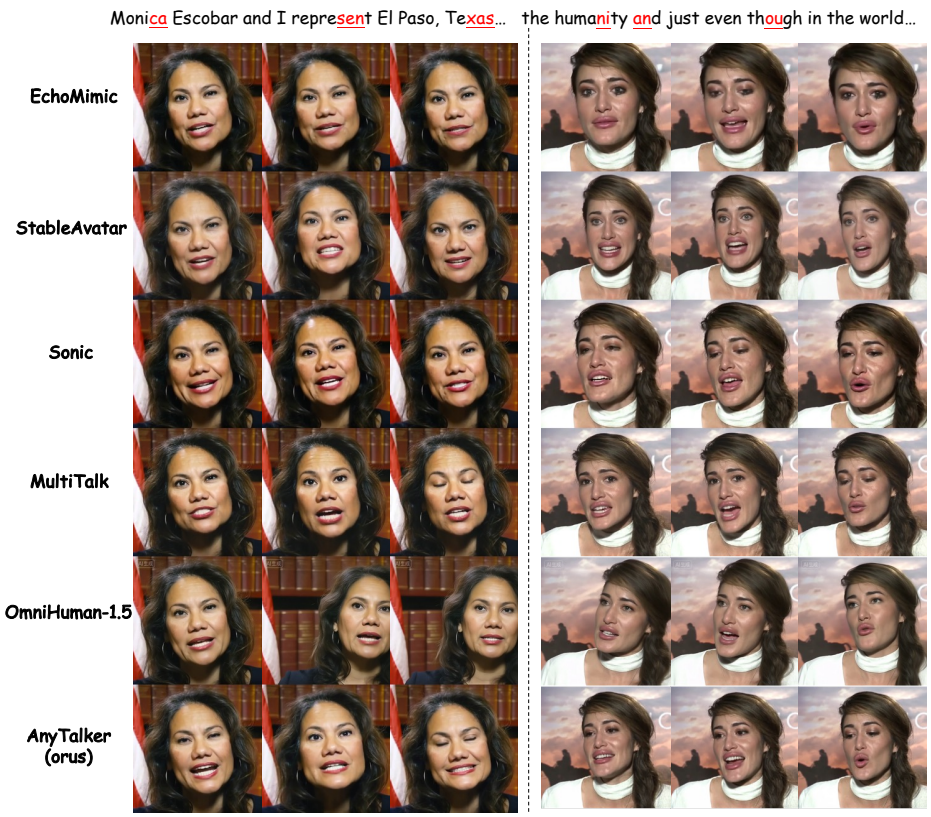


Figure 9: Qualitative results on HDTF (left) and VFHQ (right). Pronunciation positions are highlighted in red and underlined.

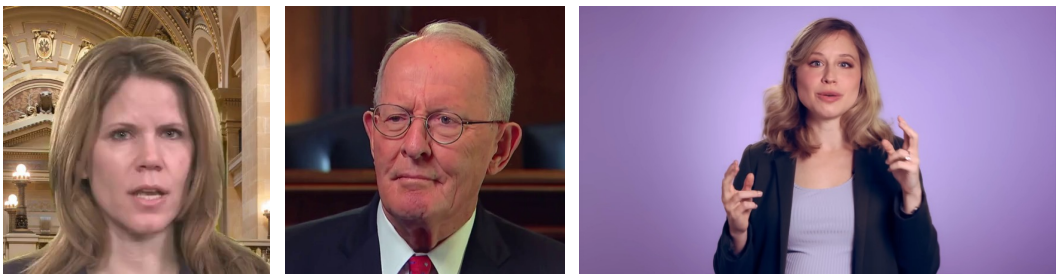


Figure 10: Benchmark reference inputs, left to right: HDTF, VFHQ, EMTD.

I Hard Case Analysis

In this section, we briefly characterize where the strategy breaks.

(i) Non-human subjects. Because the face detector used during training is unreliable for non-human subjects, we manually specify the face crop boxes at inference. We evaluate 10 animal and 10 cartoon-character inputs, each containing 2 to 4 subjects, under 100 random seeds. Approximately 30% of the non-human trials degenerate into near-static clips that repeat the first frame, compared with fewer than 5% for comparable human inputs. The detection-free backbone animates such subjects more reliably, indicating a cost of face-box supervision.

(ii) Large-scale occlusion. Prolonged full-face occlusion can alter character appearance after reappearance, a limitation shared with foundation backbones (Wan et al., 2025; Kong et al., 2024).



Figure 11: Direct qualitative comparisons under identical inputs. The upper block compares Bind-Your-Avatar, MultiTalk, Playmate2, and AnyTalker on two-person examples. The lower block compares Playmate2 and AnyTalker on three- and four-person examples because the other two baselines do not support these identity counts. Each adjacent pair shows two frames from the same generated clip.

Depth cues and memory modules are possible remedies for future work. Minor and partial overlaps, however, are handled by the global-box masking and the CLIP face cue.

(iii) Large camera motion. Stage-2 filtering removes high-motion data to stabilize training, trading away walking-while-talking scenarios. Lifting this constraint likely requires motion-robust binding supervision.

(iv) Overlapping speech. Concatenated stage-1 pairs frequently speak simultaneously. We verify overlapping-speech generation on clips with fully overlapping audio tracks.



Figure 12: Additional results generated by AnyTalker.

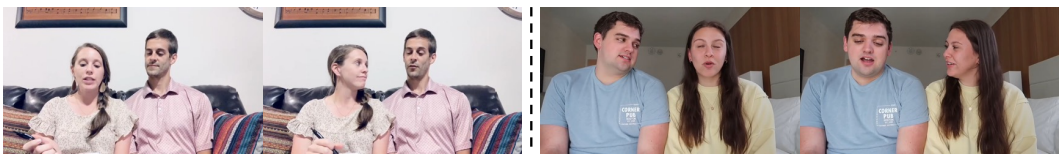


Figure 13: Two InteractivePerson cases with speaking periods for both participants.



Figure 14: Interactivity improvement after stage-2 refinement on real two-person conversation data (right).