INTERROLE: ROLE-AWARE SPEECH INJECTION FOR MULTI-CHARACTER INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-driven character animation has achieved impressive progress in single-character scenarios, but remains challenging in multi-character settings. Existing methods often rely on region-exclusive speech injection to prevent identity leakage, which improves speaker-audio binding but limits listener access to speech cues and weakens speaker-listener interactions. In this paper, we identify listener-side behavior as a central component of multi-character audio-driven animation and formulate the key thesis that speech should be globally accessible but decoded in a role-specific manner. Guided by this thesis, we analyze several speech injection schemes and propose InterRole, a complete framework for role-aware multi-character audio-driven animation. Its core module, RoleMixAttn, separates speech conditioning into expressive and receptive attention branches and uses dynamic soft routing to adaptively mix them for each video token. This design jointly models speaker articulation and listener responses while improving robustness to character movement and occlusion. To support training and evaluation, we construct the RoleTalk-AV dataset, adopt a staged training strategy, and introduce interaction-oriented metrics for gaze alignment and prosody-correlated listener motion. Experiments on Wan2.1 and LTX-2.3 show that InterRole generates more coherent speaker-listener interactions than existing multi-character baselines.

1 INTRODUCTION

The rapid advancement of video generative models and joint audio-visual foundation models, such as Wan and LTX (Wan et al., 2025; HaCohen et al., 2026) has significantly impacted audio-driven character animation. Driven by its extensive applications in gaming, filmmaking, and virtual reality, this domain has garnered substantial research attention (Fei et al., 2025; Meng et al., 2025a; He et al., 2023). Over recent years, it has achieved remarkable progress, evolving from early lip-sync generation (Chen et al., 2025b; Prajwal et al., 2020) to hand articulation (Cui et al., 2025; Shen et al., 2023) and further to full-body motion synthesis (Lin et al., 2025; Huang et al., 2025a).

Despite this progress, most existing methods use global audio cross-attention as the basic mechanism for injecting speech conditions. While effective in single-character scenarios, this design becomes fragile when extended to complex multi-character settings, where it often leads to severe identity confusion and feature leakage, such as mismatched speech-character binding or a single speech stream inadvertently driving all characters to speak. Recent studies have begun to address this issue. A common solution is to associate each speech stream with a predefined masked character region and inject the corresponding audio features only into that region (Weng et al., 2025; Huang et al., 2025b; Kong et al., 2025; Zhong et al., 2025; Wang et al., 2025c; Chen et al., 2025a). This strategy is effective for maintaining speaker-audio binding. However, it also introduces inherent limitations: 1) **Interaction Blindness**: Strictly partitioning speech into speaker regions isolates listeners from relevant speech signals, rendering listeners into static or stochastic posturing and forcing speakers to act without regard for their counterparts, ultimately stripping the animation of its necessary interactional awareness. (Figure 2, first row). While AnyTalker (Zhong et al., 2025) employs multi-person data refinement to induce listener reactions, the listener response is still not explicitly conditioned on the speaker's audio. 2) **Rigid Spatial Binding**: The reliance on predefined masks severely constrains character movement (Figure 2, second row). Although dynamic region prediction has been explored (Wang et al., 2025c), these techniques remain prone to feature leakage in complex scenarios such as character swapping or mutual occlusions.

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.

To overcome these limitations, we propose **InterRole**, built on the thesis that speech should be accessible to all characters but decoded in a role-specific manner in multi-character settings. We investigate several audio injection schemes and introduce RoleMixAttn to govern character interactions. RoleMixAttn separates speech conditioning into two functional attention branches: an expressive branch for self-driven speaking behaviors and a receptive branch for listener-side responses. This design allows the same speech representation to be decoded differently according to the role of each character. Furthermore, multi-character behavior is rarely a binary choice between speaking and listening. In real conversations, a character's motion is often a continuous mixture of self-driven expression and passive response to others. We therefore replace rigid masks with dynamic soft routing, allowing each video token to adaptively mix the expressive and receptive attention branches. This soft routing not only better captures inter-character interactions but also improves robustness to occlusion and character crossing.

We build a scalable data curation pipeline that collects, filters, and stratifies audio-visual clips into single-character, structured dialogue, and in-the-wild multi-character subsets, resulting in the RoleTalk-AV dataset. Based on this dataset, we adopt a staged training strategy. The model first learns speaker-side expression and listener-side response from large-scale single-character samples, where the expressive and receptive branches can be optimized with relatively clean role supervision. It is then refined on multi-character dialogue data to learn speaker-listener coordination under realistic spatial layouts and motion patterns. For evaluation, we complement standard metrics with two interaction-oriented proxy metrics. Gaze Saliency Alignment (GSA) measures whether listeners attend to the active speaker, while Audio Prosody Correlation (APC) measures whether listener micro-motion is correlated with speech dynamics. We instantiate InterRole on Wan2.1 for fair comparison with existing methods and further apply it to LTX-2.3 to evaluate generalization. Extensive experiments show that InterRole produces more coherent listener responses compared with prior multi-character baselines. The main contributions are summarized as follows:

- We identify listener-side behavior as a central component of multi-character animation, formulate the thesis that speech should be globally accessible but decoded in a role-specific manner, and present InterRole as a complete framework for this setting.
- We present RoleMixAttn, a role-aware audio injection module that jointly models speaker-listener interaction patterns through soft-gated expressive and receptive attention branches.
- We construct the RoleTalk-AV dataset, adopt staged training and interaction-oriented evaluation, and validate InterRole on both Wan2.1 and LTX-2.3, showing improvements over existing multi-character baselines.

2 Related Work

2.1 Audio-Driven Generation for Single Character

Speaker Head Generation. Pioneering efforts primarily focused on accurate lip-sync (Zhang et al., 2023; Cheng et al., 2022; Xu et al., 2025). Subsequent researchers began to explicitly incorporate facial pose and emotional features (Cui et al., 2024; 2025; Shen et al., 2024; Ma et al., 2023; Ji et al., 2025; Li et al., 2026). Recent advancements have expanded the scope to full-body animation (Gan et al., 2025; Gao et al., 2025; Tian et al., 2025; Zhang et al., 2025; Meng et al., 2025b; Wang et al., 2025a; Li et al., 2025). To achieve more harmonious results, (Jiang et al., 2025; Ding et al., 2025) employ MLLMs to interpret sophisticated audio semantics, and (Liang et al., 2025; Meng et al., 2025a) incorporate reinforcement learning frameworks to optimize the global coherence and quality of the generated animations. **Listener Head Generation.** Following the pioneering work of RLHG (Zhou et al., 2022), which maps speaker audio-visual signals to listener animations, subsequent works have improved motion diversity and emotional responsiveness (Ng et al., 2022; Song et al., 2023; Liu et al., 2023; Geng et al., 2023; Kong et al., 2026). More recent efforts emphasize ultra-low latency to enable seamless, real-time user interaction (Yu et al., 2025; Zhao et al., 2025; Guo et al., 2025; Ki et al., 2026; Wang et al., 2026). Despite significant advancements in these subfields, existing methods remain functionally isolated and single-character confined, and cannot resolve feature leakage and ensure interactive harmony in multi-character settings. To this end, we propose InterRole for more generalized audio-driven animation.

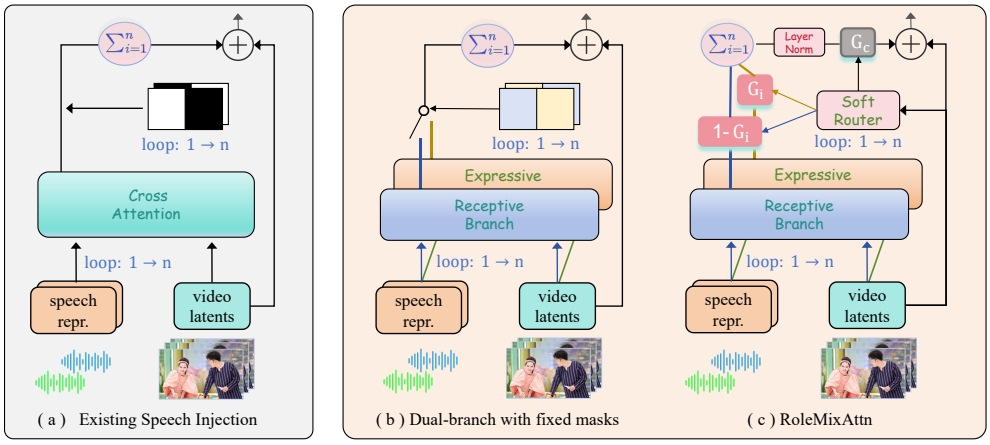


Figure 1: **Different multi-speech injection schemes.** (a) Prior methods bind each speech stream to its corresponding speaker region, which reduces leakage but isolates listeners from speech cues and weakens speaker-listener interaction. (b) We explicitly decouple speech conditioning into an expressive branch for speaker behaviors and a receptive branch for listener responses. (c) After the two branches are stabilized through staged training, we further replace binary masks with a dynamic soft router, enabling flexible mixing of expressive and receptive cues, thereby enhancing interaction awareness and improving robustness to character movement and occlusion.

2.2 Audio-Driven Multi-Character Animation

Existing research primarily focuses on mitigating feature leakage. Some methods rely on predefined masks (Wang et al., 2025b; Chen et al., 2025a). AnyTalker (Zhong et al., 2025) and Playmate2 (Ma et al., 2025) utilize masks to inject speech features into speaker regions to maintain character-speech correspondence. However, these methods often fail to accommodate character movement. To address this, some works attempt to dynamically predict character regions during the generation (Huang et al., 2025b). MultiTalk (Kong et al., 2025) leverages self-attention maps to localize characters, while InterActHuman (Wang et al., 2025c) employs a dedicated mask predictor. Despite these advancements, rigidly confining speech to the speaker's region renders listeners oblivious to vocal cues, leading to stilted and interactionally disjointed behaviors. To overcome these, we introduce InterRole, which enables all characters to perceive the global acoustic context and map audio to diverse interaction patterns.

3 Methodology

Multi-character audio-driven animation introduces a unique challenge. Given an image $\mathcal{I}$ containing multiple characters and one or more speech streams $\mathcal{A} = \{\mathbf{a}_i\}_{i=1}^{n}$, the model must animate the active speakers with accurate lip synchronization while also generating plausible listener behaviors, such as gaze shifts, nodding, and posture changes. To address this challenge, we propose InterRole, a complete framework for role-aware multi-character audio-driven animation, covering a multi-speech injection module (Section 3.1), data curation and staged training (Section 3.2), and interaction-oriented metrics (Section 3.3).

3.1 Role-Aware Multi-Speech Injection

Speech representation. For each speech stream $\mathbf{a}_i$, we extract frame-aligned speech features using a pretrained audio encoder:

$$\mathbf{x}_i = \mathrm{Enc}(\mathbf{a}_i) \in \mathbb{R}^{T \times N_a \times C_a}, \tag{1}$$

Here T, N_a, and C_a correspond to video latent frame count, per-frame audio token number, and audio feature dimension, respectively. Given video latent tokens $\mathbf{z} \in \mathbb{R}^{T \times (H \times W) \times C_v}$, we fuse $\mathbf{x}_i$ into $\mathbf{z}$ under distinct injection paradigms.

Global cross-attention. Our first attempt lets all video tokens attend to all speech streams naively:

$$\mathbf{z}_{\text{global}} = \mathbf{z} + \sum_{i=1}^{n} \text{Attn}(\mathbf{z}, \mathbf{x}_i). \tag{2}$$

where $\text{Attn}(\cdot, \cdot)$ denotes standard cross-attention. This scheme makes speech globally accessible to all characters, which is crucial for modeling listener responses. However, because it does not distinguish the role of each speech stream with respect to each character, it remains prone to speech-identity mismatch and feature leakage even with high-quality training data.

Dual-branch with fixed masks. To reduce role ambiguity, we further explore a two-branch injection scheme that explicitly decouples speaker-side and listener-side speech conditioning. Specifically, we split the cross-attention into two functional branches. The expressive branch $\mathcal{E}$ models self-driven speaking behaviors, such as lip articulation and speaking gestures, while the receptive branch $\mathcal{R}$ models listener-side responses, such as gaze shifts, nodding, and subtle facial reactions:

$$\mathbf{e}_i = \mathcal{E}(\mathbf{z}, \mathbf{x}_i), \qquad \mathbf{r}_i = \mathcal{R}(\mathbf{z}, \mathbf{x}_i). \tag{3}$$

Given a human mask $\mathbf{M}_i \in [0,1]^{(H\times W)\times 1}$ for the speech stream $\mathbf{x}_i$, this injection can be written as:

$$\mathbf{z}_{\text{fixed}} = \mathbf{z} + \sum_{i=1}^{n} \left(\mathbf{M}_i \odot \mathbf{e}_i + \mathbf{M}_{\neg i} \odot \mathbf{r}_i\right), \quad \mathbf{M}_{\neg i} = \text{clip}\left(\sum_{j\neq i} \mathbf{M}_j, 0, 1\right) \tag{4}$$

This design explicitly makes listener behavior a first-class modeling target. However, the binary mask assumption is still restrictive. In realistic conversations, a character's motion is often a continuous mixture of self-driven expression and passive response to others, which binary masks cannot represent. Moreover, binary masks are sensitive to occlusion and character movement.

RoleMixAttn. Based on the dual-branch injection, we finally introduce RoleMixAttn, which replaces binary character masks with dynamic soft routing. Each speech stream $\mathbf{x}_i$ is bound to a designated character through a user-supplied bounding box. We use this matched box to crop the corresponding identity image $\mathcal{I}_i$ from the input image $\mathcal{I}$ and encode $\mathcal{I}_i$ using the VAE encoder shared with the base video model, obtaining identity-aware visual tokens $\mathbf{v}_i^{\text{id}}$. A lightweight multi-head cross-attention module then aggregates these tokens with learnable queries, followed by linear projection, to produce a compact identity representation $\mathbf{q}_i$. Conditioned on $\mathbf{q}_i$, RoleMixAttn predicts a token-wise soft routing score for each speech stream $\mathbf{x}_i$:

$$\mathbf{G}_i = \sigma\left(\Psi\left(\mathbf{z}, \mathbf{q}_i\right)\right), \qquad \mathbf{G}_i \in [0,1]^{T\times(H\times W)\times 1}, \tag{5}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and $\Psi(\cdot)$ is a two-layer MLP. A high value in $\mathbf{G}_i$ indicates that the token is more likely to correspond to the speaker of $\mathbf{x}_i$, so the expressive branch dominates. A low value indicates that the token belongs to another character, so the receptive branch dominates. The speech-conditioned update is then computed as:

$$\Delta\mathbf{z} = \sum_{i=1}^{n} \left(\mathbf{G}_i \odot \mathcal{E}(\mathbf{z}, \mathbf{x}_i) + (1 - \mathbf{G}_i) \odot \mathcal{R}(\mathbf{z}, \mathbf{x}_i)\right). \tag{6}$$

These soft routing values allow the model to adaptively combine both branches, capturing interaction patterns that cannot be represented by rigid masks. Since $\mathbf{G}_i$ is learned and updated dynamically from the current video latents, the model is also more robust to character occlusion and movement.

To suppress speech-induced artifacts in background regions, we introduce a character gate $\mathbf{G}_{\text{c}}$. This gate estimates whether each video token belongs to a character by computing the semantic relevance between the video latents $\mathbf{z}$ and a learnable character anchor $\mathcal{C}$:

$$\mathbf{G}_{\text{c}} = \sigma\left(\Phi\left(\mathbf{z}, \mathcal{C}\right)\right), \qquad \mathbf{G}_{\text{c}} \in [0,1]^{T\times(H\times W)\times 1}, \tag{7}$$

where $\Phi(\cdot)$ is a two-layer MLP. The final output of RoleMixAttn is:

$$\mathbf{z}_{\text{out}} = \mathbf{z} + \mathbf{G}_{\text{c}} \odot \mathbf{LayerNorm}(\Delta\mathbf{z}). \tag{8}$$

As $\Delta\mathbf{z}$ sums contributions from all n speech streams, LayerNorm normalizes its scale before the residual addition, keeping the update numerically stable as the number of characters grows.

3.2 ROLETALK-AV AND STAGED TRAINING

High-quality multi-character dialogue data is scarce, while single-character videos and noisy multi-person videos are much easier to obtain. We therefore construct RoleTalk-AV, a curated audio-visual dataset comprising heterogeneous data types. We collect candidate clips from publicly available sources. Low-quality samples are removed according to visual resolution, audio quality, shot stability, face visibility, and audio-visual synchronization. We then apply ByteTrack (Zhang et al., 2022) to detect and track all visible faces across each clip, obtaining both the number of concurrent characters and their spatial motion trajectories. Based on the tracked face count and motion magnitude, the data is stratified into three subsets:

$$\mathcal{D} = \{\mathcal{D}_{\text{single}}, \mathcal{D}_{\text{dialogue}}, \mathcal{D}_{\text{wild}}\}. \tag{9}$$

$\mathcal{D}_{\text{single}}$ (1K hours) consists of single-person shots featuring either a standalone speaker or a listener reacting to off-screen audio. $\mathcal{D}_{\text{dialogue}}$ (10 hours) contains structured multi-character dialogues with relatively stable layouts and clear speaker-listener roles. $\mathcal{D}_{\text{wild}}$ (80 hours) contains in-the-wild multi-character clips with larger motion, occlusion, viewpoint changes, and more diverse interaction patterns. We then use SAM-Audio 3 (Shi et al., 2025) to separate the mixed soundtrack into individual speech streams and employ TalkNet (Tao et al., 2021) to estimate correspondence between speeches and characters. More details on data collection, filtering, role assignment, quality control, and statistics are provided in the supplement. The dataset and model weights will be publicly released upon acceptance, and the source code is included in the supplement.

To exploit the heterogeneous structure of RoleTalk-AV and the decoupled design of our architecture, we adopt a staged training strategy.

Stage 1: role-specific branch pretraining. We first train the expressive and receptive branches on $\mathcal{D}_{\text{single}}$ while bypassing the gating module. We optimize only the expressive branch for speaker samples to learn speaking expressions and solely update the receptive branch for listener samples to model speech-conditioned non-verbal reactions. This stage delivers clear role supervision and stabilizes both branches before full multi-character training, with the standard diffusion loss $\mathcal{L}_{\text{diff}}$ used for optimization.

Stage 2: structured dialogue training. In this stage, we use tracked character regions to warm-start the routing gates in RoleMixAttn. Since the gates are randomly initialized, directly learning soft routing in multi-character scenes can cause ambiguous branch assignment and destabilize early training. The structured dialogue subset contains relatively reliable speaker-listener layouts and character tracks, making it suitable for gate supervision. We use these tracks only as an initialization signal: $\mathbf{G}_i$ is encouraged to localize the character associated with speech stream $\mathbf{x}_i$, while $\mathbf{G}_{\text{c}}$ suppresses non-character background regions. We define the foreground character mask as:

$$\mathbf{M}_{\text{c}} = \text{clip}\left(\sum_i \mathbf{M}_i, 0, 1\right), \tag{10}$$

where $\mathbf{M}_i$ denotes the tracked region of the i-th person. The gate supervision loss is:

$$\mathcal{L}_{\text{gate}} = \text{BCE}(\mathbf{G}_{\text{c}}, \mathbf{M}_{\text{c}}) + \sum_i \text{BCE}(\mathbf{G}_i, \mathbf{M}_i). \tag{11}$$

The overall objective is:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{gate}}\mathcal{L}_{\text{gate}}, \tag{12}$$

where λ_{gate} controls the supervision strength. This warm start stabilizes convergence and provides a structured initialization for the following training phase.

Stage 3: in-the-wild refinement. Finally, we refine the full model on $\mathcal{D}_{\text{wild}}$ to improve robustness under realistic multi-character conditions, including character crossing, mutual occlusion, and dynamic camera motion. In this stage, we remove the explicit gate supervision and optimize the model only with $\mathcal{L}_{\text{diff}}$. This allows the routing module to move beyond noisy tracking masks and learn more flexible and effective routing mechanisms for modeling multi-character interaction patterns in in-the-wild scenes.

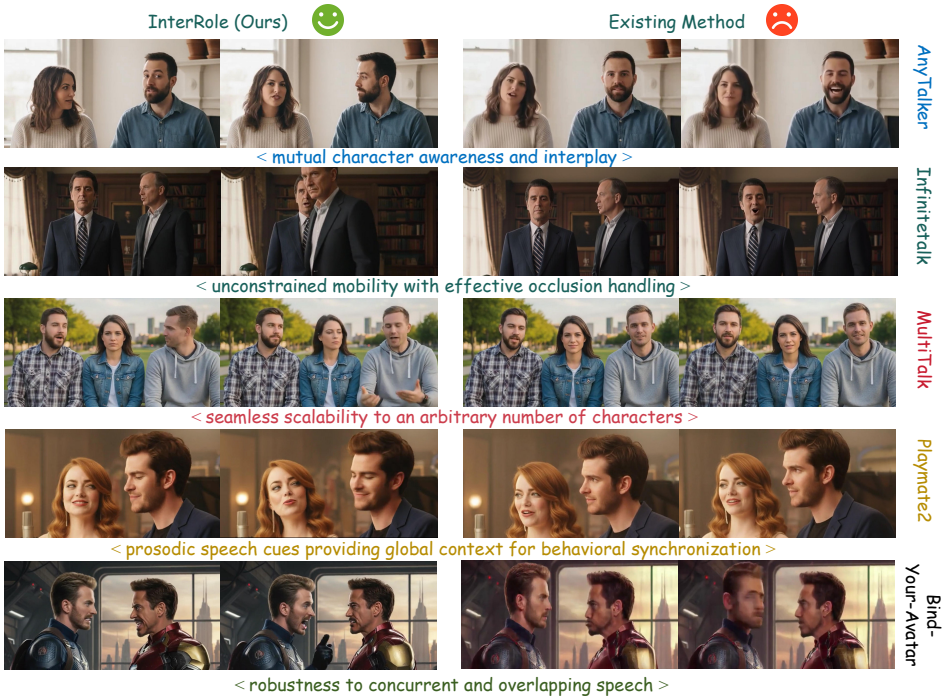


Figure 2: **Qualitative Comparison.** Our InterRole framework supports a diverse range of advanced capabilities inherently.

3.3 Interaction-Oriented Metrics

Standard evaluation metrics mainly focus on visual quality and lip synchronization. These metrics are necessary but insufficient for multi-character animation, where listener behavior and speaker-listener coordination are also important. We therefore complement standard metrics with two interaction-oriented proxy metrics.

GSA. Gaze Saliency Alignment measures whether a listener attends to the active speaker. Let H_t denote the active speaker's head region at frame t, and let G_t denote the listener's gaze target estimated by a gaze estimator (Ryan et al., 2025). For a sequence of length T, GSA is defined as:

$$\mathrm{GSA} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}(G_t \in H_t). \tag{13}$$

where $\mathbb{1}(\cdot)$ is the indicator that outputs 1 if the condition is met and 0 otherwise. A higher GSA indicates that the listener's gaze is more frequently aligned with the speaker.

APC. Audio Prosody Correlation measures whether listener micro-motion is correlated with speech dynamics. Let E_t denote the speech energy envelope and M_t denote the listener's non-mouth facial and head motion. To account for natural response delay, APC is computed as the maximum normalized cross-correlation over a temporal lag τ:

$$\mathrm{APC} = \max_{\tau} \frac{\sum_t (E_t - \bar{E})(M_{t+\tau} - \bar{M})}{\sqrt{\sum_t (E_t - \bar{E})^2 \sum_t (M_{t+\tau} - \bar{M})^2}}. \tag{14}$$

We further validate these two proxy metrics by measuring their Pearson correlation with human-rated interaction quality. GSA and APC achieve correlations of 0.67 and 0.74, a level comparable to that of widely used automatic proxy metrics, supporting their use as auxiliary indicators that complement human evaluation.

Table 1: **Quantitative Comparison.** InterRole shows consistent improvements over existing multi-character audio-driven animation methods.

Method	FID ↓	FVD ↓	Sync-C ↑	GSA ↑	APC ↑
MultiTalk	40.19	<u>392.98</u>	7.04	0.52	<u>0.37</u>
Bind-Your-Avatar	54.87	573.70	4.51	0.33	0.27
Playmate2	<u>33.16</u>	399.23	6.98	0.47	0.34
InfiniteTalk	36.51	401.29	<u>7.08</u>	0.51	0.36
AnyTalker-14B	42.46	425.38	7.07	<u>0.54</u>	0.32
InterRole	**31.26**	**358.79**	**7.12**	**0.59**	**0.46**

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

We use Wav2Vec 2.0 (Baevski et al., 2020) to extract the speech representation for each separated speech stream. Since most prior multi-character audio-driven animation methods are built on Wan2.1-I2V-14B (Wan et al., 2025), we adopt it as our main generative backbone for fair comparison. RoleMixAttn is inserted into every transformer block and λ_{gate} is set to 1. Each block has an independent routing gate, and the expressive and receptive branches are configured with 40 and 20 attention heads, respectively. We also adapt InterRole to the recent LTX-2.3 model (HaCohen et al., 2026). Specifically, we replace LTX-2.3's original audio cross-attention with RoleMixAttn: the original audio attention is used to initialize the expressive branch, while a newly receptive branch and soft routing module are introduced to model listener-side speech responses. For both backbones, we freeze the base model and optimize only the audio injection modules. The model is trained with the AdamW optimizer (Loshchilov & Hutter, 2017) using a learning rate of 2×10^{-5} on 16 NVIDIA H100 GPUs.

4.2 EVALUATION DETAILS

We evaluate generation quality from both speaker-side fidelity and speaker-listener coordination. For visual quality and temporal coherence, we use FID and FVD. For active speakers, we report Sync-C. For listener behavior, we use the proposed GSA and APC metrics. We compare InterRole against current representative methods capable of audio-driven multi-character synthesis, encompassing Bind-Your-Avatar (Huang et al., 2025b), MultiTalk (Kong et al., 2025), Playmate2 (Ma et al., 2025), InfiniteTalk (Yang et al., 2025), and AnyTalker-14B (Zhong et al., 2025). Given the absence of a public benchmark, we curate a comprehensive test set following the collection protocol of MultiTalk (Kong et al., 2025). We further enrich the test set with challenging scenarios, including speakers moving while talking and mutual occlusion. The final test set contains 300 curated samples, all carefully deduplicated from the training data. **We additionally provide video demonstrations in the supplement, where side-by-side comparisons illustrate the advantages of InterRole in producing coherent speaker-listener interactions.**

4.3 COMPARISON

Except for Bind-Your-Avatar, all compared methods are built upon Wan2.1-I2V-14B. To provide a more rigorous comparison and disentangle the effect of the proposed architecture from that of the training data, we fine-tune all baselines on RoleTalk-AV under their author-recommended configurations. At test time, we keep all input and inference conditions identical across methods, including the text prompt, speech conditions, reference image, sampling strategy, and other generation hyperparameters.

4.3.1 QUANTITATIVE COMPARISON.

As shown in Table 1, InterRole achieves the best performance across all generation-quality metrics. The improvements in FID and FVD indicate better visual quality and temporal coherence, while the

Table 2: **User study.** We report the preference rate (%) of InterRole over representative baselines.

Baseline	Binding	Listener Resp.	Interaction
MultiTalk	60.2	72.8	70.4
InfiniteTalk	54.7	74.9	73.5
AnyTalker-14B	52.5	67.6	66.8

gain in Sync-C shows that the proposed role-aware injection does not sacrifice speaker lip synchronization. More importantly, InterRole obtains clear improvements on GSA and APC, suggesting that listeners not only attend more consistently to the active speaker, but also exhibit motion patterns that are more correlated with speech prosody. These results support our central claim that multi-character audio-driven animation requires not only accurate speaker-audio binding, but also explicit modeling of listener-side speech responses.

Furthermore, comparative experiments in the supplement demonstrate InterRole's versatility and efficiency in single-character settings by selectively activating only the expressive branch ($\mathcal{E}$).

4.3.2 Qualitative Comparison.

The results are shown in Figure 2. In the first row, our result illustrates that the characters naturally alternate their gaze between the counterpart and the camera, with the listener providing contextually appropriate responses to the speech prosody, whereas the baseline tends to exhibit static motion. In the second row, our model enables free character movement and maintains robust audio-character correspondence under severe occlusion, while the baseline fails to follow movement prompts and remains confined to fixed positions. In the third row, our approach supports three-party interactions, surpassing the baseline's two-person limit. The fourth row shows that our model can use prosodic speech cues to coordinate collective behavioral patterns, such as rhythmic head-nodding, which are not captured by prior methods. The fifth row further shows that our model maintains robustness in concurrent and overlapping speech scenarios. These observations show that InterRole improves the coherence and interaction awareness of audio-driven multi-character animation.

4.3.3 User Study.

We conduct a pairwise user study on 300 randomly sampled cases from the held-out test set. We compare InterRole with representative baselines using the same settings. For each pair, 100 participants are shown two anonymized videos in random order and asked to choose the better result, with ties allowed, in terms of speaker-audio binding, listener responsiveness, and overall interaction coherence. As shown in Table 2, InterRole is consistently preferred over existing methods, with the largest gains observed in listener responsiveness and overall interaction coherence.

Additionally, participants provide a 5-point interaction quality score for each video. We compute Pearson correlation coefficients between the averaged human scores and the proposed metrics (GSA and APC) to evaluate their consistency with human perception. GSA and APC achieve Pearson correlations of 0.67 and 0.74, respectively, indicating positive alignment with human judgments.

4.4 Ablation Study

4.4.1 Ablation of Module Design.

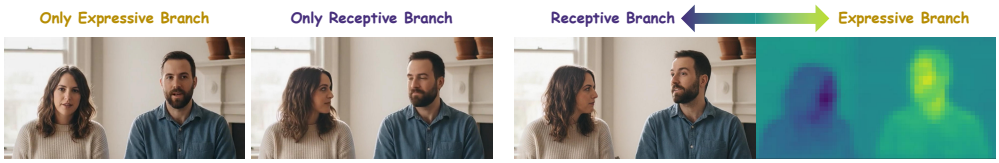
Table 3 compares different speech injection and routing designs. Global audio cross-attention performs poorly because it cannot distinguish the role of each character for a given speech stream, leading to audio-identity mismatch and feature leakage. Separating expressive and receptive branches with fixed masks substantially improves performance by explicitly modeling speaker and listener behaviors. Learned binary routing supports dynamic spatial assignment but remains restrictive, as each token must select only one branch. In contrast, the proposed soft routing adaptively combines both branches and achieves the best overall generation quality and interaction performance.

Figure 3 (a) further verifies the functional specialization of the two branches. Routing all tokens through the expressive branch causes both characters to exhibit speaker-like articulation, whereas receptive-only routing produces listener-like responses. Under the male character's speech input,

Table 3: **Ablation Study.** We evaluate the effects of speech injection schemes, routing strategies, staged training, and backbone adaptation on generation quality and speaker-listener coordination.

Setting	FID ↓	FVD ↓	Sync-C ↑	GSA ↑	APC ↑
Global audio cross-attention	48.71	468.35	5.86	0.39	0.29
Dual-branch w/ fixed mask	<u>37.21</u>	405.27	7.04	0.48	0.34
Dual-branch w/ binary route	39.84	<u>386.64</u>	**7.14**	0.46	0.37
w/o Staged training	38.50	394.18	6.87	0.40	0.35
w/o Character gate	39.75	401.26	7.09	<u>0.54</u>	<u>0.42</u>
InterRole	**31.26**	**358.79**	<u>7.12</u>	**0.59**	**0.46**
LTX-2.3-finetune	37.80	375.40	6.79	0.42	0.40
LTX-2.3-InterRole	**32.92**	**362.76**	**7.03**	**0.58**	**0.48**

Figure 3 (b) visualizes the corresponding branch-routing weights G_{male}. The gate maps preserve speech-character correspondence while exhibiting non-binary values, indicating that the expressive and receptive branches cooperate to generate context-dependent motion. Finally, removing the character gate degrades performance (Table 3), confirming its role in suppressing speech-induced background artifacts.



(a) Visualization of single-branch forced routing (b) Visualization of the gating values

Figure 3: **RoleMixAttn branch analysis.** (a) Forced single-branch routing experiments for branch role validation. (b) Gating weight heatmaps of G_{male} with male speech input.

4.4.2 Ablation of Training Strategy.

We ablate our staged training pipeline by directly training full RoleMixAttn on mixed data. Jointly optimizing expressive and receptive branches from scratch hinders each branch from learning distinct functional roles, yielding inferior lip synchronization and unstable listener motions. Our staged scheme instead first builds role priors on clean single-person clips, then learns speaker-listener coordination via structured dialogues, and finally boosts generalization on wild multi-character footage. This progressive training consistently lifts generation quality and interaction performance.

4.4.3 Backbone Generalization.

We evaluate whether the proposed injection strategy generalizes to a stronger audio-video backbone. We compare direct fine-tuning of LTX-2.3's original audio cross-attention with replacing it with RoleMixAttn in Table 3. The RoleMixAttn variant achieves better speaker-listener coordination, demonstrating that the benefit of our method does not come from a specific Wan2.1 implementation, but from the role-aware speech injection design itself.

5 Conclusion

We presented InterRole, a role-aware framework for multi-character audio-driven animation. Motivated by the need to model listener behavior in multi-character interactions, InterRole makes speech globally accessible while decoding it in a role-specific manner. Its RoleMixAttn module separates expressive and receptive speech conditioning and adaptively mixes them with dynamic soft routing. Together with RoleTalk-AV, staged training, and interaction-oriented evaluation, InterRole improves speaker-listener coordination and overall animation quality across different video backbones.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, including improving the grammar, clarity, and concision of the manuscript. We did not use generative AI to formulate the core research ideas, design the proposed method, generate or analyze experimental results, or select citations; the remaining disclosure categories are not applicable. All AI-assisted text was reviewed and revised by the authors. We manually verified the technical claims, equations, citations, code, and experimental results. We take full responsibility for the final content of this work, including any text or artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

Our work involves the generation of realistic human animations and may therefore be misused to create misleading or impersonating content. We intend this work solely for research and legitimate creative applications and encourage users to clearly disclose generated content and obtain appropriate consent when depicting identifiable individuals. RoleTalk-AV is constructed from publicly available videos. We exclude private or sensitive content and content involving minors when identifiable, and we release only metadata, derived annotations, and download scripts when permitted by the licenses of the original sources. We do not claim ownership of the source videos. We acknowledge that demographic and geographic biases in online video data may be inherited by the dataset and model, and we encourage further evaluation of these limitations before deployment.

REPRODUCIBILITY STATEMENT

We provide the model implementation in the supplementary material. The main paper describes the RoleMixAttn architecture, staged training strategy, objective functions, evaluation metrics, implementation settings, and baseline protocols. The appendix provides additional details on dataset construction and quality control, architectural comparisons, computational analysis, and supplementary experiments. Together, these materials are intended to support reproduction of the proposed method and evaluation results. We plan to publicly release the code and model weights upon acceptance, subject to applicable licensing and responsible-use requirements.

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*, 2020.

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*, 2025b.

Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia 2022 Conference Papers*, 2022.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, 2016.

Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*, 2024.

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*, 2025.

Yikang Ding, Jiwen Liu, Wenyuan Zhang, Zekun Wang, Wentao Hu, Liyuan Cui, Mingming Lao, Yingchao Shao, Hui Liu, Xiaohan Li, et al. Kling-avatar: Grounding multimodal instructions for cascaded long-duration avatar animation synthesis. *arXiv preprint arXiv:2509.09595*, 2025.

Zhengcong Fei, Hao Jiang, Di Qiu, Baoxuan Gu, Youqiang Zhang, Jiahua Wang, Jialin Bai, Debang Li, Mingyuan Fan, Guibin Chen, et al. Skyreels-audio: Omni audio-conditioned talking portraits in video diffusion transformers. *arXiv preprint arXiv:2506.00830*, 2025.

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.

Scott Geng, Revant Teotia, Purva Tendulkar, Sachit Menon, and Carl Vondrick. Affective faces for goal-driven dyadic communication. *arXiv preprint arXiv:2301.10939*, 2023.

Ying Guo, Xi Liu, Cheng Zhen, Pengfei Yan, and Xiaoming Wei. Arig: Autoregressive interactive head generation for real-time conversations. *arXiv preprint arXiv:2507.00472*, 2025.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Tianyu He, Junliang Guo, Runyi Yu, Yuchi Wang, Jialiang Zhu, Kaikai An, Leyi Li, Xu Tan, Chunyu Wang, Han Hu, et al. Gaia: Zero-shot talking avatar generation. *arXiv preprint arXiv:2311.15230*, 2023.

Donglin Huang, Yongyuan Li, Tianhang Liu, Junming Huang, Xiaoda Yang, Chi Wang, and Weiwei Xu. Vividanimator: An end-to-end audio and pose-driven half-body human animation framework. *arXiv preprint arXiv:2510.10269*, 2025a.

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*, 2025b.

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*, 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.

Taekyung Ki, Sangwon Jang, Jaehyeong Jo, Jaehong Yoon, and Sung Ju Hwang. Avatar forcing: Real-time interactive head avatar generation for natural conversation. *arXiv preprint arXiv:2601.00664*, 2026.

Xiangyu Kong, Xiaoyu Jin, Yihan Pan, Haoqin Sun, Hengde Zhu, Xiaoming Xu, Xiaoming Wei, Lu Liu, and Siyang Song. Echo: Towards emotionally appropriate and contextually aware interactive head generation. *arXiv preprint arXiv:2603.17427*, 2026.

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.

Hebeizi Li, Benyuan Sun, Yi Yang, Zihao Liang, Zihao Yin, Xiao Sha, and Chenliang Wang. Unitalking: A unified audio-video framework for talking portrait generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4647–4656, 2026.

Xiaodi Li, Pan Xie, Yi Ren, Qijun Gan, Chen Zhang, Fangyuan Kong, Xiang Yin, Bingyue Peng, and Zehuan Yuan. Infinityhuman: Towards long-term audio-driven human. *arXiv preprint arXiv:2508.20210*, 2025.

Chao Liang, Jianwen Jiang, Wang Liao, Jiaqi Yang, Weihong Zeng, Han Liang, et al. Alignhuman: Improving motion and fidelity via timestep-segment preference optimization for audio-driven human animation. *arXiv preprint arXiv:2506.11144*, 2025.

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*, 2025.

Jin Liu, Xi Wang, Xiaomeng Fu, Yesheng Chai, Cai Yu, Jiao Dai, and Jizhong Han. Mfr-net: Multi-faceted responsive listening head generation via denoising diffusion model. In *Proceedings of the 31st ACM international conference on multimedia*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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.

Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. Dreamtalk: When emotional talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*, 2023.

Rang Meng, Yan Wang, Weipeng Wu, Ruobing Zheng, Yuming Li, and Chenguang Ma. Echomimicv3: 1.3 b parameters are all you need for unified multi-modal and multi-task human animation. *arXiv preprint arXiv:2507.03905*, 2025a.

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*, 2025b.

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

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 *Proceedings of the 28th ACM international conference on multimedia*, 2020.

Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M Rehg. Gaze-lle: Gaze target estimation via large-scale learned encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *CVPR*, 2023.

Xiaoqian Shen, Faizan Farooq Khan, and Mohamed Elhoseiny. Emotalker: Audio driven emotion aware talking head generation. In *Proceedings of the Asian Conference on Computer Vision*, 2024.

Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, et al. Sam audio: Segment anything in audio. *arXiv preprint arXiv:2512.18099*, 2025.

Luchuan Song, Guojun Yin, Zhenchao Jin, Xiaoyi Dong, and Chenliang Xu. Emotional listener portrait: Realistic listener motion simulation in conversation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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

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*, 2025a.

Qiang Wang, Mengchao Wang, Fan Jiang, Yaqi Fan, Yonggang Qi, and Mu Xu. Fantasyportrait: Enhancing multi-character portrait animation with expression-augmented diffusion transformers. *arXiv preprint arXiv:2507.12956*, 2025b.

Yinuo Wang, Yanbo Fan, Xuan Wang, Boyao Zhou, Yu Guo, Yujun Shen, and Fei Wang. Mimictalker: A multimodal interactive and memory-enhanced framework for real-time dyadic 3d head generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32399–32409, 2026.

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.

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

Shuchen Weng, Haojie Zheng, Zheng Chang, Si Li, Boxin Shi, and Xinlong Wang. Audio-sync video generation with multi-stream temporal control. *arXiv preprint arXiv:2506.08003*, 2025.

Zhihua Xu, Tianshui Chen, Zhijing Yang, Siyuan Peng, Keze Wang, and Liang Lin. Exploiting temporal audio-visual correlation embedding for audio-driven one-shot talking head animation. *IEEE Transactions on Multimedia*, 2025.

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*, 2025.

Haojie Yu, Zhaonian Wang, Yihan Pan, Meng Cheng, Hao Yang, Chao Wang, Tao Xie, Xiaoming Xu, Xiaoming Wei, and Xunliang Cai. Llia–enabling low-latency interactive avatars: Real-time audio-driven portrait video generation with diffusion models. *arXiv preprint arXiv:2506.05806*, 2025.

Jiangning Zhang, Junwei Zhu, Zhenye Gan, Donghao Luo, Chuming Lin, Feifan Xu, Xu Peng, Jianlong Hu, Yuansen Liu, Yijia Hong, et al. Soul: Breathe life into digital human for high-fidelity long-term multimodal animation. *arXiv preprint arXiv:2512.13495*, 2025.

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

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, 2022.

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*, 2021.

Sizhe Zhao, Chenyang Wang, Weiyu Zhao, Zonglin Li, Ming Li, and Shengping Zhang. Realistener: Real-time listening head generation with dynamic emotion modeling and flexible modality adaptation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

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

Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. Responsive listening head generation: a benchmark dataset and baseline. In *European conference on computer vision*, 2022.

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, 2022.

A Architectural Property Comparison

We first analyze the architectural properties of existing multi-character audio-driven animation methods. Specifically, we consider three properties: (1) *listener audio access*, which indicates whether non-speaking characters can directly access the active speech signal for listener-side response generation; (2) *dynamic spatial support*, which indicates whether the method can handle large character movement, crossing, and occlusion without relying on fixed speaker regions; and (3) *multi-character scalability*, which indicates whether the formulation naturally extends beyond a fixed two-character setting.

Table 4: **Architectural property comparison.** The $\surd$ and $\times$ indicate whether each method satisfies the corresponding architectural prerequisite according to its published design.

Method	Listener Audio Access	Dynamic Spatial Support	Multi-Character Scalability
MultiTalk	$\times$	$\surd$	$\times$
Bind-Your-Avatar	$\times$	$\surd$	$\surd$
Playmate2	$\times$	$\times$	$\surd$
InfiniteTalk	$\times$	$\surd$	$\times$
AnyTalker-14B	$\times$	$\times$	$\surd$
InterRole	$\surd$	$\surd$	$\surd$

As shown in Table 4, most prior multi-character methods rely on localized speech injection. This design is effective for reducing speaker-audio leakage, but it also prevents listeners from directly accessing speech cues, making listener behavior weakly conditioned or largely stochastic. In addition, methods based on predefined masks, such as AnyTalker and Playmate2, are naturally constrained by fixed spatial assignments and therefore struggle with large motion, character crossing, and mutual occlusion. Methods with hard-coded two-person spatial encodings, such as MultiTalk and InfiniteTalk, further limit scalability to scenes involving more than two characters. In contrast, InterRole makes speech globally accessible while decoding it through role-specific expressive and receptive branches. Together with dynamic soft routing, this design supports listener-side speech conditioning, flexible character movement, and extension to multiple characters within a unified formulation.

B RoleTalk-AV Dataset

To support role-aware multi-character audio-driven animation, we construct **RoleTalk-AV**, a large-scale audio-visual dataset containing both single-character and multi-character conversational videos. Unlike existing talking-head or multi-person video datasets that primarily emphasize active speaker articulation, RoleTalk-AV is designed to capture both speaker-side expressive behaviors and listener-side receptive responses, including gaze shifts, nodding, facial reactions, posture changes, and turn-taking dynamics.

Data sources. We collect candidate videos from publicly available online sources, including interviews, talk shows, podcasts, panel discussions, debates, lectures, and in-the-wild conversational clips. All videos are processed only when they satisfy the following basic requirements: visible human characters, audible speech, sufficient visual resolution, and temporal continuity. We remove videos with severe compression artifacts, excessive camera cuts, strong background noise, or poor audio-visual synchronization. The collected videos are used solely for research purposes, and we will release metadata, processing scripts, and derived annotations according to the license constraints of the original sources.

Preprocessing. Each raw video is first resampled to a unified frame rate of 16 fps and segmented into clips of 5–10 seconds. We then apply an automatic quality filtering pipeline. Specifically, we discard clips with low visual resolution, low face visibility, unstable shot transitions, severe motion blur, heavy occlusion, or poor speech quality. Audio-visual synchronization is estimated using an off-the-shelf synchronization model (Chung & Zisserman, 2016), and clips with synchronization confidence below a threshold are removed. To avoid data leakage, near-duplicate clips are identified

Table 5: **Statistics of RoleTalk-AV.** We report the duration, number of clips, average number of visible characters, and primary usage of each subset.

Subset	Hours	Clips	Avg. chars	Usage
RoleTalk-Single	1000	590K	1.0	Branch pretraining
RoleTalk-Dialogue	10	5.2K	2.2	Gate warm-up
RoleTalk-Wild	80	47K	2.7	Robust refinement
Total	1090	642K	1.1	Full training

using perceptual video hashes and audio fingerprints, and duplicates are removed before train/test splitting.

Character tracking and role estimation. For each retained clip, we detect and track visible characters across frames. Face regions are detected using ByteTrack (Zhang et al., 2022), producing frame-level character tracks and bounding boxes. We then use SAM-Audio 3 (Shi et al., 2025) to separate the mixed soundtrack into individual speech streams. We further estimate active speaker correspondence using an audio-visual active speaker detection model, TalkNet (Tao et al., 2021). Given a separated or mixed speech stream, the model assigns speaker confidence scores to each visible character at each frame. A character is labeled as an active speaker when the confidence exceeds a threshold and remains temporally consistent for at least 4 frames. Non-speaking visible characters during active speech intervals are treated as listener candidates. Clips with ambiguous speaker assignment, severe overlap between speakers, or unstable tracks are either discarded or assigned to the in-the-wild subset with lower-confidence annotations.

Speech stream processing. We also estimate speech energy, pitch, and prosodic features for downstream analysis of listener motion. Clips with unreliable separation or inconsistent speaker-track assignment are filtered out during quality control.

Dataset stratification. RoleTalk-AV is stratified into three subsets according to the number of visible characters, role structure, and scene complexity:

- **RoleTalk-Single** contains single-character clips, including active speaker clips and listener-like clips where the visible character reacts to off-screen speech. This subset is mainly used to learn clean role-specific motion priors for expressive and receptive behaviors.
- **RoleTalk-Dialogue** contains structured multi-character dialogue clips with relatively stable layouts, clear speaker-listener roles, and reliable character tracks. This subset is used to warm-start role routing and learn speaker-listener coordination under controlled multi-character settings.
- **RoleTalk-Wild** contains challenging in-the-wild multi-character clips with camera motion, character crossing, partial occlusion, viewpoint changes, turn-taking, and occasional overlapping speech. This subset is used to improve robustness under realistic multi-character interactions.

Dataset statistics. Table 5 summarizes the statistics of RoleTalk-AV. The dataset contains 1090 hours of video in total, including 1K hours of single-character clips, 10 hours of structured dialogue clips, and 80 hours of in-the-wild multi-character clips. It covers a wide range of conversational settings.

Quality control. We combine automatic filtering with manual inspection to ensure annotation quality. For each subset, we randomly sample 500 clips and manually verify face tracking, active speaker assignment, speech separation quality, and speaker-listener role labels. The estimated annotation precision is 95.6% for active speaker assignment and 97.3% for character tracking. We also manually inspect difficult cases such as occlusion, character crossing, and overlapping speech to estimate the failure rate of the automatic pipeline. Clips with unreliable annotations are excluded from RoleTalk-Dialogue and retained only in RoleTalk-Wild when they still provide useful in-the-wild motion diversity.

Privacy and ethics. RoleTalk-AV is constructed from publicly available videos and is intended only for research on audio-driven animation. We avoid collecting private or sensitive content and remove clips involving minors, private conversations, or personally sensitive situations when identifiable. Since the dataset may contain real human identities, we will release only the processed metadata, annotations, and download scripts when permitted by source licenses. We also encourage future users to follow responsible-use guidelines and avoid generating misleading or impersonating content.

C SINGLE-CHARACTER COMPARISON

To further demonstrate the versatility of our framework, we evaluate its performance on standard talking-head benchmarks, HDTF (Zhang et al., 2021) and CelebV-HQ (Zhu et al., 2022). We compare InterRole with current single-character state-of-the-art methods, including EchoMimic (Chen et al., 2025b), Hallo3 (Cui et al., 2025), FantasyTalking (Wang et al., 2025a), AniPortrait (Wei et al., 2024), Sonic (Ji et al., 2025), and OmniAvatar (Gan et al., 2025).

Table 6: Quantitative comparison with other state-of-the-art single-character methods on HDTF and CelebV-HQ benchmarks.

Method	**HDTF**			**CelebV-HQ**		
	Sync-C↑	**FID↓**	**FVD↓**	**Sync-C↑**	**FID↓**	**FVD↓**
EchoMimic	5.36	60.82	240.07	4.16	63.72	318.08
Hallo3	6.55	33.98	153.31	5.57	40.81	**212.91**
FantasyTalking	3.61	32.64	103.01	3.14	37.54	218.43
AniPortrait	3.09	32.83	112.21	2.09	<u>37.17</u>	250.24
Sonic	<u>8.35</u>	29.53	**89.34**	<u>6.68</u>	39.89	224.48
OmniAvatar	7.06	**26.63**	112.49	5.88	37.41	250.67
InterRole	**8.94**	<u>28.44</u>	<u>101.25</u>	**7.12**	**35.15**	<u>217.57</u>

Notably, we only load and activate weights related to the expressive space during inference, thereby eliminating unnecessary complexity and computational overhead. Quantitative results, as detailed in Table 6, demonstrate that our method consistently achieves state-of-the-art or competitive performance across all metrics on single-character benchmarks.

Table 7: Comparison of computational efficiency for generating a 49-frame video on an NVIDIA H100 GPU.

Method	**MultiTalk**	**Playmate2**	**InterRole**
Inference Time (min) ↓	3.50	12.52	3.61
Peak GPU Memory (GB) ↓	56.82	51.68	56.10

D LIMITATIONS

While InterRole establishes a new paradigm for interactive multi-character animation, several challenges remain for future exploration: 1) Although our method theoretically supports an unlimited number of characters, the current implementation requires sequential processing of each speech stream $\mathbf{x}_i$. Consequently, inference latency scales linearly with the number of active speakers, which may pose challenges for applications involving large groups. 2) Our framework currently focuses on generating temporally continuous video. However, professional multi-character productions (e.g., films or talk shows) frequently utilize shot transitions to emphasize specific speakers. Enhancing our training datasets with such cinematic cuts represents a promising direction for extending our model's stylistic versatility in future work. 3) Consistent with the limitations of existing video diffusion backbones, the model may encounter quality degradation in extremely long-duration sequences.

E COMPUTATIONAL ANALYSIS

To empirically validate the efficiency, we benchmark InterRole against two representative baselines, MultiTalk (Kong et al., 2025) and Playmate2 (Ma et al., 2025). All three models are built upon the same foundational backbone, Wan2.1-I2V-14B (Wan et al., 2025).

Our profiling on an NVIDIA H100 GPU for generating a 49-frame 480P video shows that InterRole maintains a highly competitive computational profile, as detailed in Table 7. Specifically, both our average inference time and peak GPU memory usage remain closely aligned with the baselines. These results empirically validate that our specialized modules provide significant architectural advantages with minimal computational and memory overhead, ensuring scalability for complex multi-character synthesis.