

## Technical Appendices and Supplementary Material

In this appendix, we provide details of the model architecture in Appendix B, training setting in Appendix C, data processing pipeline in Appendix D, and we provide comprehensive ablation studies and experiment results in Appendix A. We discuss limitations and broader impact in Appendix E and Appendix F

## A Experiments

### A.1 Audio Cross-Attention Design Ablation

To better understand how to inject audio conditioning into DiT-based video diffusion models, we conduct a systematic design study comparing multiple audio cross-attention mechanisms. Due to the high computational cost of training the full 30B MoCha model, we perform our ablation using a smaller 4B DiT backbone pretrained on MovieGen [8], and fine-tune it on 400K close-up, single-character clips with audio annotations for 200K training steps.

All models share identical training hyperparameters and differ only in the design of the audio cross-attention module. We evaluate lip-sync quality on MoCha-Bench using the SyncNet-based metrics Sync-C ( $\uparrow$ ) and Sync-D ( $\downarrow$ ). Results are summarized in Table 4.

**Variants.** Below, we define the attention query  $q \in \mathbb{R}^{N \times d}$  as the video tokens, and key/value  $k, v \in \mathbb{R}^{T \times d}$  as the audio tokens. We denote  $A(q, k, v)$  as the standard attention function.

- **Naive Audio Cross-Attention:** Each video token attends to all audio tokens without positional encoding:

$$z = A(q, k, v).$$

- **+ Learnable Positional Embedding:** Audio tokens are augmented with learnable positional embeddings  $p_j$  initialized to zero:

$$z = A(q, k + p, v).$$

- **+ Sinusoidal Positional Embedding:** Instead of learned  $p_j$ , we add fixed sinusoidal embeddings:

$$z = A(q, k + \text{Sinusoidal}(j), v).$$

- **+ Rotary Positional Embedding (RoPE):** RoPE is applied by rotating the queries and keys:

$$z = A(\text{RoPE}(q), \text{RoPE}(k), v).$$

- **Localized Audio Attention (proposed):** Each video token  $q_i$  attends only to a window of audio tokens  $k_{[j_0:j_1]}$  centered on its corresponding temporal segment (see Section 3.2):

$$z_i = A(q_i, k_{[j_0:j_1]}, v_{[j_0:j_1]}),$$

where  $j_0 = \max(1, (i-1)r-1)$  and  $j_1 = \min(T, ir+1)$ .

- **Localized Audio Attention + RoPE:** We additionally apply RoPE to  $q$  and  $k$  within the window.

**Findings.** As shown in Table 4, the naive attention baselines perform poorly in lip-sync accuracy. Adding positional information—especially sinusoidal or RoPE embeddings—significantly improves performance, suggesting that positional priors are critical for learning speech-video alignment. However, our proposed **Localized Audio Attention** consistently outperforms all other variants, demonstrating the effectiveness of constrained temporal windows for resolving the resolution mismatch between video and audio tokens.

Interestingly, adding RoPE to the Localized Attention variant slightly degrades performance, indicating potential interference between the inductive bias introduced by RoPE and the explicit temporal alignment imposed by windowing.

Method	Sync-C $\uparrow$	Sync-D $\downarrow$
DiT-4B + Localized Audio Attention ( <b>MoCha-4B</b> )	<b>5.692</b>	<b>8.403</b>
DiT-4B + Localized Audio Attention + RoPE	5.027	9.038
DiT-4B + Naive Audio Attention + RoPE	4.872	8.893
DiT-4B + Naive Audio Attention + Sinusoidal Embedding	4.747	8.986
DiT-4B + Naive Audio Attention + Learnable Embedding	2.540	10.363
DiT-4B + Naive Audio Attention	2.364	10.385

Table 4: **Comparison of Audio Cross-Attention Variants on MoCha-Bench.** We report lip-sync metrics. The proposed Localized Audio Attention achieves the best performance.

## A.2 Ablation of Curriculum-Based Multimodal Training Strategy

We conduct an ablation study to assess the effectiveness of our proposed *Curriculum-based Multimodal Training* strategy (described in Section 3.4), which is designed to address two core challenges: (i) data scarcity and limited diversity in speech-annotated datasets, and (ii) varying speech relevance across spatial scales in different shot types.

To evaluate this strategy, we compare the full MoCha-30B model with two ablated variants:

- **w/o Curriculum Training:** Trained on mixed modalities (speech and text-only) but without curriculum progression (i.e., trained directly on mixed data with full shot-type complexity from the beginning).
- **w/o Mixed Multimodal Training:** Trained solely on speech-annotated videos, without any text-only data or curriculum scheduling.

All models use the same architecture and are trained for an equal number of steps.

We report both automatic lip-sync metrics (Sync-C  $\uparrow$ , Sync-D  $\downarrow$ ) and three human evaluation metrics from MoCha-Bench: *Text Alignment*, *Visual Quality*, and *Action Naturalness* (detailed in Section A.4). Qualitative examples can be found on the anonymous website.

As shown in Table 5, removing the curriculum phase (*w/o Curriculum Training*) causes a moderate performance drop across all metrics, confirming the benefit of staged training that gradually increases visual complexity. Although trained for the same number of steps, the non-curriculum baseline converges more slowly and underperforms.

The *w/o Mixed Multimodal Training* variant—trained only on limited speech-annotated data—performs significantly worse, especially on *Text Alignment* and *Action Naturalness*. This confirms that unimodal speech-driven training causes overfitting to front-facing talking-face data, impairing the model’s ability to generalize to diverse prompts for diverse scenes and full-body activities.

These results validate the necessity of both **modality mixing** and **curriculum progression** for robust and generalizable talking character generation.

Method	Sync-C $\uparrow$	Sync-D $\downarrow$	Text Alignment $\uparrow$	Visual Quality $\uparrow$	Action Naturalness $\uparrow$
<b>MoCha-30B (with Curriculum)</b>	<b>6.037</b>	<b>8.103</b>	<b>3.85</b>	<b>3.72</b>	<b>3.82</b>
w/o Curriculum Training	5.659	8.435	3.17	3.31	3.27
w/o Mixed Multimodal Training	5.798	8.231	2.71	2.91	2.97

Table 5: **Ablation of Curriculum-Based Multimodal Training Strategy on MoCha-Bench.** We report lip-sync metrics (Sync-C  $\uparrow$ , Sync-D  $\downarrow$ ) and human evaluation scores across three axes.

## A.3 Ablation of Character Tagging Strategy

We ablate the effectiveness of the *Character Tagging* strategy introduced in Section 4 for handling multi-character conversations with turn-based dialogue. This strategy is particularly important in multi-clip scenes, where multiple characters appear across different segments and speaker transitions are inferred from the audio input alone.

Our tagging mechanism assigns a unique identifier (e.g., Person1, Person2) to each character introduced at the beginning of the prompt. These tags are then reused across individual clip descriptions, allowing for clear, consistent references without repeating verbose appearance descriptions. This structured prompting significantly reduces prompt length and improves character consistency (see Figure 5).

To evaluate the impact of character tagging, we compare the full MoCha-30B model (with tagging) to a baseline trained with naïve captioning—where detailed character descriptions are repeated verbosely in each clip. Both models are evaluated on the turn-based dialogue subset of MoCha-Bench.

As shown in Table 6, removing character tagging results in a drastic drop in *Text Alignment*, indicating the model often confuses which character appears in which scene. Qualitative examples on the anonymous website show that, without tagging, the model may generate scene-swap artifacts, fail to transition characters correctly between clips, or maintain similar lip-sync for mismatched dialogue.

We also observe degradation in *Visual Quality*, as the model occasionally blends inconsistent character features across clips. However, lip-sync metrics remain relatively stable, as the audio continues to guide temporal speech alignment. Overall, these results highlight that character tagging is essential for multi-character consistency and semantic alignment in dialogue-driven video generation.

Method	Sync-C $\uparrow$	Sync-D $\downarrow$	Text Alignment $\uparrow$	Visual Quality $\uparrow$	Action Naturalness $\uparrow$
<b>MoCha-30B (with Character Tagging)</b>	<b>5.432</b>	<b>8.461</b>	<b>3.81</b>	<b>3.64</b>	<b>3.69</b>
w/o Character Tagging	5.486	8.465	2.01	2.15	3.03

**Table 6: Ablation of Character Tagging on MoCha-Bench (Turn-Based Dialogue Category).** We report lip-sync metrics (Sync-C  $\uparrow$ , Sync-D  $\downarrow$ ) and human evaluation scores. Character tagging improves semantic consistency and reduces scene confusion in multi-clip dialogue scenarios.

#### A.4 MoCha-Bench Human Evaluation

We conduct a comprehensive human evaluation to compare MoCha against baseline methods on the MoCha-Bench dataset. The evaluation is based on five axes tailored for the Talking Characters task, with scores ranging from 1 to 4. Each model output received 5 independent ratings per example, resulting in over 1000 responses per model. We provide the evaluation guidance as below. Besides the text guideline, we also include some visual examples to better help the annotators to judge.

This document provides evaluation guidelines, including axis definitions, scoring rubrics, and instructions for annotators. Visual examples are provided separately to support consistent judgments.

#### Task Overview

Each evaluation sample consists of:

- A generated video with audio,
- A text prompt describing the scene and character behavior.

Your task is to evaluate how well the generated video across five dimensions.

#### Evaluation Axes

- **Lip-Sync Quality:** Measures how accurately the character’s lip movements align with the spoken audio.  
*Scale:* 1 – Not aligned at all, 2 – Weak alignment, 3 – Mostly aligned, 4 – Perfectly aligned.
- **Facial Expression Naturalness:** Evaluates whether facial expressions and lip-sync appear natural and contextually appropriate, without seeming robotic or exaggerated.  
*Scale:* 1 – Completely unnatural, 2 – Noticeably synthetic or stiff, 3 – Mostly natural and believable, 4 – Indistinguishable from real or cinematic performance.
- **Action Naturalness:** Assesses how naturally the character’s body movements and gestures align with the audio.  
*Scale:* 1 – Completely unnatural, 2 – Noticeably unnatural, 3 – Mostly natural, 4 – Indistinguishable from real movie or TV characters.

- **Text Alignment:** Measures how well the generated actions, expressions, and presence of characters follow the behaviours described in the prompt.  
*Scale:* 1 – No alignment (e.g., missing character or major misbehavior),  
2 – Partial alignment, 3 – Mostly aligned, 4 – Perfect alignment with the prompt.
- **Visual Quality:** Evaluates the overall visual fidelity, including image sharpness, coherence, and absence of rendering issues such as artifacts, glitches, or anatomical distortions (e.g., broken limbs or unnatural body proportions).  
*Scale:* 1 – Severe artifacts, 2 – Noticeable artifacts, 3 – Mostly artifact-free, 4 – Flawless visuals.

## Annotation Instructions

1. **Review the prompt and watch the full video at least once.**
2. **Evaluate each axis independently.** Use the provided 1–4 scale. Do not assign the same score across all axes unless truly justified.
3. **Use the full scale range.** Assign low or high scores as needed. Avoid defaulting to the midpoint.
4. **If no character is present in the video, assign a score of 1 for *Text Alignment*.**

## FAQ

### Q: What if the same scene includes multiple characters?

A: *Focus your evaluation on the central or speaking character as described in the prompt. Other characters may be present, but your ratings should reflect the behavior and alignment of the primary subject.*

### Q: What if the video does not include a character?

A: *If the generated video fails to include the main character described in the prompt (e.g., an empty scene or wrong subject), you should assign a score of 1 for *Text Alignment*.*

Method	Lip-Sync Quality	Facial Expression Naturalness	Action Naturalness	Text Alignment	Visual Quality
Hallo3 [17]	<u>2.45</u>	<u>2.25</u>	<u>2.13</u>	<u>2.35</u>	<u>2.36</u>
SadTalker [28]	1.21	1.14	1.00	N/A	2.95
AniPortrait [29]	1.16	1.12	1.00	N/A	1.45
<b>MoCha (Ours)</b>	<b>3.85 (+1.40)</b>	<b>3.82 (+1.57)</b>	<b>3.82 (+1.69)</b>	<b>3.85 (+1.50)</b>	<b>3.72 (+1.36)</b>

Table 7: **Human evaluation scores on MoCha-Bench.** Scores range from 1 to 4 across five evaluation axes, where a score of 4 reflects performance that is nearly indistinguishable from real video or cinematic production. Participants rated each method on five aspects: lip-sync quality, facial expression naturalness, action naturalness, text-prompt alignment, and visual quality. MoCha significantly outperforms prior methods across all categories. Green numbers indicate absolute improvements ( $\Delta$ ) over the second-best method (underlined). SadTalker and AniPortrait consistently received a score of 1 for action naturalness, as these methods only perform head movements.

## A.5 MoCha-Bench Qualitative Comparison

Figure 8 presents a direct comparison between MoCha and baseline methods on MoCha. All baselines require a reference image as an auxiliary input. To ensure fairness, we first generate a video using MoCha and then use its first frame as the reference image for all baseline models. For models that do not support arbitrary aspect ratios, we crop the first frame to focus on the head region before feeding it into their networks. We provide two groups of qualitative comparisons: one featuring close-up shots and the other medium shots. The close-up group emphasizes lip-sync quality, head movement, and facial expressions, while the medium shot group focuses on hand movements during speech. MoCha not only produces lip movements that closely align with the input speech—enhancing both articulation and naturalness—but also generates expressive facial animations and realistic, coordinated actions that accurately follow the textual prompt. In contrast, SadTalker and AniPortrait exhibit minimal

head motion and limited lip synchronization. While Hallo3 achieves mostly consistent lip-syncing, it suffers from inaccurate articulation and erratic head movements. In the medium shot comparisons, Hallo3 also introduces noticeable visual artifacts, particularly during complex actions.

## B MoCha Model Architecture

### B.1 3D VAE

Our 3D-VAE is based on a variational autoencoder and compresses the input pixel space video  $\nu$  of shape  $\in \mathbb{R}^{T \times H \times W \times 3}$  into a lower-dimensional, continuous latent representation  $x_0$  of shape  $\tau \times h \times w \times c$ . In our implementation, we compress the input  $8\times$  across the spatial dimension while  $4\times$  across the temporal dimension, i.e.,  $H/h = W/w = 8$  and  $T/\tau = 4$ . The latent channel dimensionality is fixed at  $c = 16$  for all experiments reported herein.

### B.2 Text Encoder

We employ a triad of text encoding architectures—UL2, ByT5, and a Long-prompt variant of MetaCLIP—to equip the backbone with both high-level semantic and fine-grained character-level textual comprehension. Each encoder generates a sequence of 256 token embeddings. To unify these heterogeneous representations, we apply dedicated linear projection layers and LayerNorm to each encoder’s output, transforming them into the model dimension 6144-dimensional feature space. The resulting normalized embeddings from all three streams are then concatenated to produce the final comprehensive text representation fed into the backbone. Among these, the MetaCLIP encoder specializes in generating text features inherently aligned with visual modalities, optimizing performance for cross-modal generation tasks. UL2, conversely, excels at encoding deep linguistic reasoning and semantics, while ByT5 captures character-level details, making it effective for encoding visual text.

### B.3 Audio Encoder

Our audio pipeline is powered by Wav2Vec2, but instead of relying solely on its final output, we extract and stitch together the embeddings from all 12 internal layers. This approach gives us a deeper, layered view of the audio content, with each layer contributing a 768-dimensional slice to the overall representation. After running the audio through Wav2Vec2’s tokenizer, we stretch or compress the resulting sequence using linear interpolation before the audio hits Wav2Vec2. So that we end up with the same number of audio features as there are video frames—effectively assigning a unique audio token to each frame. To provide each frame’s audio token with extra context, we expand its feature vector by gluing on the tokens from the five frames before and after it. So for any given frame  $f$ , the final embedding is built as  $A(f) = [A(f-5), \dots, A(f), \dots, A(f+5)]$ . This chunky, context-aware audio feature then passes through a straightforward two-layer neural net (an MLP with a hidden size of 512) which reshapes it into the 6144-dimensional token  $\alpha(f)$  needed by our model’s backbone.

### B.4 DiT Architecture

The core architectural hyperparameters of our MoCha-30B DiT backbone are provided in Table 8.

**Localized Audio Cross-Attention.** To incorporate temporal locality in the audio stream, we adopt a windowed cross-attention mechanism with a window size of  $r+2=6$ , where  $r=4$  is the temporal downsampling factor of the video encoder. For a latent video frame  $x^{(i)} \in \mathbb{R}^{h \times w \times c}$  at timestep  $i \in \{1, \dots, \tau\}$ , attention is restricted to audio tokens  $\alpha^{(j)}$  falling within the interval:

$$j \in [\max(1, (i-1)r-1), \min(T, ir+1)], \quad (5)$$

which corresponds to the original (pre-downsampled) temporal span of  $x^{(i)}$  with one token of padding on each side to enable smoother context transitions at boundaries.

**Modality Integration Pathway.** Each attention block integrates modalities in a sequential manner. It begins with video self-attention, followed by localized audio cross-attention and then text cross-

Layers	Model Dimension	FFN Dimension	Attention Heads	Activation Function	Normalization
48	6144	16384	48	SwiGLU	RMSNorm

Table 8: Core architecture hyperparameters for the MoCha-30B Transformer. The model has 30 billion parameters in the Transformer stack alone, excluding auxiliary components such as text embedding models, speech embedding models, and the 3D-VAE.

attention. Each stage incorporates a residual connection, modulated by a scalar weight. In our implementation, these residual weights are fixed to 1. The full process is given by:

$$(1) \quad \mathbf{z}_{\text{video}}^{\text{out}} = \text{SelfAttn}_{\text{video}}(\mathbf{z}_{\text{in}})$$

$$(2) \quad \mathbf{z}_{\text{text}}^{\text{out}} = \text{CrossAttn}_{\text{text}}(\mathbf{z}_{\text{video}}^{\text{out}}, \mathbf{y}) + \lambda_{\text{text}} \cdot \mathbf{z}_{\text{video}}^{\text{out}}$$

$$(3) \quad \mathbf{z}_{\text{audio}}^{\text{out}} = \text{CrossAttn}_{\text{audio}}(\mathbf{z}_{\text{text}}^{\text{out}}, \boldsymbol{\alpha}) + \lambda_{\text{audio}} \cdot \mathbf{z}_{\text{text}}^{\text{out}}$$

where:

- $\mathbf{z}_{\text{in}}$  is the input latent sequence.
- $\text{SelfAttn}_{\text{video}}$  denotes self-attention over video tokens.
- $\text{CrossAttn}_{\text{audio}}(\cdot, \boldsymbol{\alpha})$  performs localized cross-attention with audio tokens  $\boldsymbol{\alpha}$ .
- $\text{CrossAttn}_{\text{text}}(\cdot, \mathbf{y})$  performs cross-attention with global text tokens  $\mathbf{y}$ .
- $\lambda_{\text{audio}}$  and  $\lambda_{\text{text}}$  are residual weights, both set to 1 in our implementation.

This staged fusion enables progressive enrichment of the video representation by incorporating auditory and textual information, while preserving intermediate features through residual addition.

## C MoCha Training

We provide the training details for our MoCha-30B model in Table 9. We used a constant learning rate scheduler with 2000 warm-up steps. We use a Progressive Curriculum for Multimodal Training as describe in subsection 3.4.

**Mixed-Modal Sampling.** Our data consists of a balanced mix of multimodal and unimodal data:

- 80% Multimodal (speech+text): The majority enables fine-grained audiovisual grounding.
- 20% Unimodal (text-only): Text-only video samples offer broader visual diversity and varied camera movements, helping the model retain strong generalization capabilities. In this setting, the speech embedding is replaced with a zero vector before the audio projector.

**Shot-Type-Based Curriculum.** We organize training into multiple stages based on shot complexity:

- Stage 0: We pretrain on large-scale text-only datasets to establish strong visual priors.
- Stages 1–N: We begin with close-up shots, which have high speech-visual correlation, and progressively incorporate more challenging scenarios such as medium/wide shots and multi-character scenes. At each stage, we halve the share of easier examples from the previous stage while maintaining the 80%/20% multimodal-unimodal ratio.

This combined strategy allows MoCha to benefit from both abundant text-only data and limited multimodal supervision while progressively mastering harder generation tasks.

We build MoCha based on Movie Gen Backbone. The Stage 0 training is included in the Movie Gen backbone pertaining. Then we add the speech cross attention and speech projector to the Movie Gen backbone to build MoCha. During Stages 1–N training, We full-finetuning the entire 30B MoCha model while freezing the text encoder and speech encoder and text projector. Throughout training, all

input examples are resized to a resolution of approximately 720 px, preserving their original aspect ratios.

Stage	#GPUs	Global bs	LR	Shot Types	#Iters	#Audio Samples	#Text Samples
0	1024	512	1e-5	Images / Videos	200K	100M	–
1	512	512	1e-5	Close-Up (1 Char)	200K	400k	100M
2	512	512	1e-5	Close-Up / Medium Close (1 Char)	200K	300k	100M
3	512	512	1e-5	Close-Up / Medium / Medium Close (1 Char)	100K	200k	100M
4	512	512	1e-5	Close / Medium / Multi-Char / Multi-Clip	100K	100k	100M

Table 9: Progressive Curriculum for Multimodal Training in MoCha. Stages gradually increase in temporal and compositional complexity, progressing from static images and short clips to multi-shot, multi-character, dialogue-driven video. Residual text supervision is maintained throughout.

## D Data Processing Pipeline

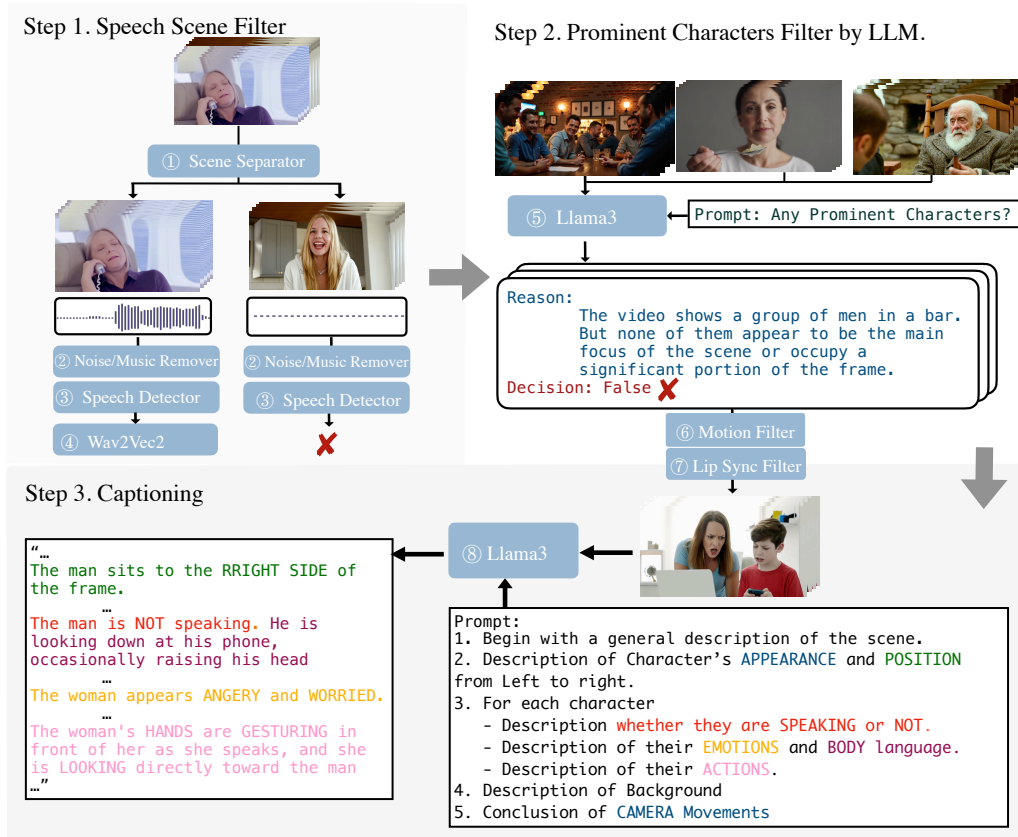


Figure 9: **Data Processing Pipeline.** Our four-stage pipeline includes: (1) **Speech Scene Filtering**, (2) **Prominent Character Filtering**, (3) **Motion and Lip-Sync Verification**, and (4) **Scene Captioning**. Together, these steps produce high-quality, speech-aligned training samples.

To ensure high-quality supervision for speech-conditioned video generation, we develop a four-stage data processing pipeline, as illustrated in Figure 9. Each stage is designed to filter noisy data and produce rich, structured annotations for training.

- **(1) Speech Scene Filtering:** Raw videos are first segmented into scenes using PySceneDetect [50]. We then detect and retain segments containing clean, spoken audio by removing clips with excessive background noise or dominant music. For valid segments, we apply noise suppression and extract speech embeddings using Wav2Vec2 [51, 23].

- **(2) Prominent Character Filtering:** To focus on scenes with clearly visible speakers, we use an LLM-based filter that analyzes visual cues and removes clips lacking a central character figure. This step emphasizes narrative relevance and visual clarity.
- **(3) Motion and Lip-Sync Verification:** We further refine the data by checking for facial and body motion aligned with speech. Segments without meaningful movement or with weak audio-visual correspondence are excluded.
- **(4) Scene Captioning:** Finally, a large language model [52] is used to generate structured captions for each clip. These include detailed descriptions of character appearance, spatial layout, speaking behavior, emotional expression, and physical gestures—enabling rich conditioning during training.

This pipeline results in a curated dataset of approximately 300 hours of high-quality, speech-aligned video content, totaling around 800K annotated samples.

## E Limitations

Despite the strong performance of our model across various talking character scenarios, we identify several limitations that affect generation quality under certain conditions. We provide corresponding examples on the [anonymous website](#).

*Failure to Lip Sync in Wide or Extreme Wide Shots:* When the input caption is too vague—particularly lacking details about facial attributes or shot type—the model may default to generating a wide or extreme wide shot. In such cases, the character often appears too far from the camera, and the lip region contains only a few pixels. As a result, the model may fail to perform accurate lip synchronization. *Example prompt: “A man playing skateboard at a skatepark.”*

*Multiple Characters in Scene:* While the model is generally capable of making the intended character speak in scenes with multiple characters, we observed occasional confusion about which character should be speaking once both characters’ face are visible in a single shot. This can lead to degraded lip-sync quality. The limitation likely stems from a scarcity of similar multi-character examples in the training data, where two or more characters appear in the same shot and speak to the camera. *Example prompt: “A medium shot set in a dimly lit tavern. The central figure, a rugged man with long, wet-looking hair and a thick beard, sits on a rustic wooden bench. He wears a weathered wool cloak over leather armor, evoking the image of a battle-hardened warrior. His expression is intense as he speaks, holding a short sword in his right hand. To his left, another man with tied-back hair and fur-lined garments watches him closely.”*

*Over-Expression with High Speech CFG:* Increasing the speech classifier-free guidance (CFG) value beyond the default (e.g., from 7.5 to 12) can cause the model to generate characters with overly expressive facial and body motions. While this can improve speech emphasis, it may also reduce realism or break immersion in otherwise grounded scenes. *Example prompt: “A tracking shot circling around the man as he ties a tie over his blue suit. He speaks to the camera while adjusting the knot, maintaining eye contact throughout.”*

## F Broader Impact

The goal of MoCha is to advance the field of dialogue-centric video generation and enable creative professionals—such as filmmakers, animators, educators, and content creators—to produce emotionally engaging character-driven videos using natural language and speech inputs. By lowering the barrier to cinematic-quality storytelling, this technology democratizes digital media production and unlocks new possibilities for interactive entertainment, educational content, and virtual communication.

However, as with any powerful generative technology, there are important societal and ethical considerations to address.

**Misuse and Synthetic Media Risks.** The use of speech-driven character generation raises concerns around the potential misuse of synthetic media. While MoCha synthesizes characters and scenes entirely from noise, without cloning real human identities or voices, there is still a risk that the generated videos could be misrepresented as real footage—particularly if aligned with real-world audio. This raises issues of misinformation, media authenticity, and potential psychological manipulation.



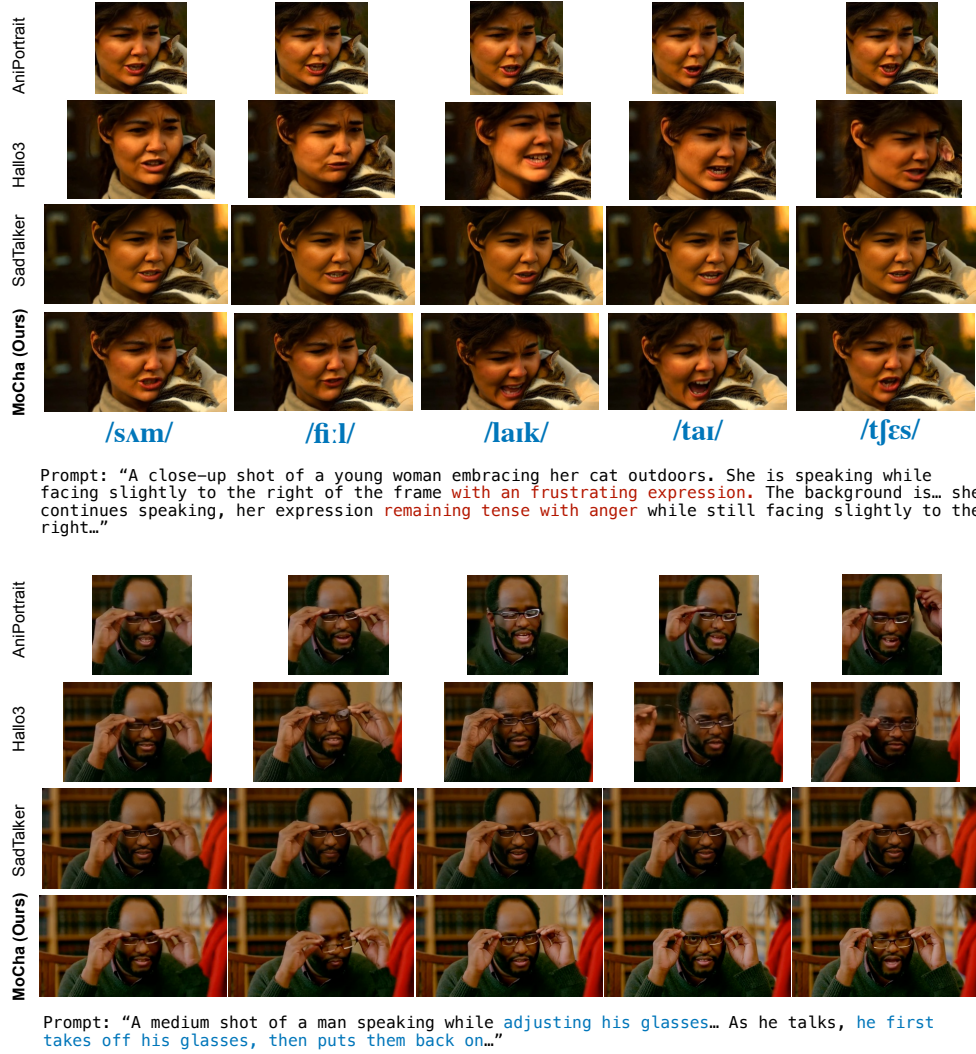


Figure 10: **Qualitative comparison between MoCha and baselines on MoCha-Bench.** MoCha not only produces lip movements that align closely with the input speech—enhancing the clarity and naturalness of articulation—but also generates expressive facial animations and realistic, complex actions that faithfully follow the textual prompt. In contrast, SadTalker and AniPortrait exhibit minimal head motion and limited lip synchronization. Hallo3 mostly follows the lip-syncing but suffers from inaccurate articulation, erratic head movements, and noticeable visual artifacts. Since the baselines operate in an image-to-video (I2V) setting, we provide them with the first frame generated by MoCha as input for comparison. The first frame is cropped and resized as needed to meet the requirements of each baseline.

1040 Unlike deepfake systems that typically manipulate real people’s faces, MoCha does not operate on  
 1041 real identity inputs. All visual content is generated from scratch using text and speech prompts.  
 1042 Nevertheless, the audio guidance—if paired with sensitive or impersonated speech—could still be  
 1043 used to create misleading portrayals. To mitigate this risk, we recommend responsible disclosure  
 1044 practices and support for watermarking or synthetic media detection tools.

1045 **Privacy and Consent.** Although MoCha does not require real human images or voices for generation,  
 1046 broader deployment of similar technologies in the future may prompt privacy concerns, especially if  
 1047 adapted to personalized avatars or voice-based likenesses. To ensure responsible use, it is critical  
 1048 to uphold strict data privacy standards, including transparent data usage policies, informed consent  
 1049 when training on human likenesses, and tools for individuals to opt out of potential misuse.



Prompt: "A close-up shot of a man sitting on a dark gray couch... Behind the man are three white cylindrical light fixtures with yellow lights inside them... the man continues to speak to the camera while he holds a lit cigar, the smoke curling gently into the air..."



Prompt: "A medium shot of a young man aged 25 to 35 is sitting in the living room in a leisurely environment... He is live-streaming, sitting in front of a desk with a laptop in front of him. His demeanor is relaxed and friendly, gesturing with his hands while speaking..."



Prompt: "A close-up shot of a young blonde woman sitting in an airplane seat, facing slightly to the right as she talks on the phone with a worried expression. As the video progresses, she continues speaking and eventually turns to look out the window..."

Figure 11: **Qualitative results of MoCha on MoCha-Bench.** MoCha not only generates lip movements that are well-synchronized with the input speech, but also produces natural facial expressions that reflect the prompt along with realistic hand gestures and action movements

1050 **Bias and Representation.** As with many generative models, outputs from MoCha may reflect biases  
 1051 present in the underlying training data—particularly in terms of character appearance, behavior, or  
 1052 cultural representation. Careful dataset curation and evaluation are needed to ensure diversity and  
 1053 inclusiveness, and to avoid reinforcing harmful stereotypes or excluding underrepresented groups in  
 1054 generated media.

1055 **Ethical Deployment.** We encourage deployment of MoCha only in contexts that respect consent,  
 1056 truthfulness, and creative integrity. As the technology matures, we advocate for the establishment of  
 1057 community-driven ethical guidelines, collaboration with media regulators, and public education on  
 1058 the capabilities and limits of AI-generated video.

1059 By proactively identifying and addressing these risks, we aim to support the safe and beneficial  
 1060 advancement of talking character technologies and ensure they are developed and applied in ways  
 1061 that are ethical, inclusive, and socially responsible.