PLANTALK: MLLM-BASED MOTION PLANNING WITH HIERARCHICAL AUDIO GUIDANCE FOR TALKING HEAD GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-driven talking head generation is challenging because facial articulation is identity-dependent and fine-grained audio-visual correspondence is difficult to preserve across multi-stage generation. In this paper, we propose **PlanTalk** to exploit joint speech-identity context through an MLLM-based motion planner, together with hierarchical audio guidance. The motion planner jointly interprets the driving speech and speaker-specific visual characteristics to organize an identity-aware facial motion sequence. Hierarchical audio guidance then progressively strengthens audio-visual correspondence at different representation levels: audio-landmark fusion refines the alignment between facial geometry and speech articulation, while denoising-stage audio attention preserves speech-consistent facial dynamics during video synthesis. The proposed cross-modal planning strategy effectively captures identity-dependent articulation, while hierarchical audio guidance maintains fine-grained speech correspondence throughout the generation pipeline. Extensive experiments demonstrate that PlanTalk achieves state-of-the-art performance over existing methods in terms of lip synchronization and identity consistency while maintaining high visual fidelity.

1 INTRODUCTION

Audio-driven talking head generation Zhou et al. (2021); Zhang et al. (2023b); Wang et al. (2021a); Ji et al. (2021) aims to synthesize facial videos synchronized with driving speech, enabling applications such as film production Prajwal et al. (2020), video conferencing Wang et al. (2021b), and digital avatars He et al. (2024). Unlike generic portrait animation, talking head generation requires the generated facial dynamics to maintain precise correspondence with the temporal and phonetic structure of speech, in addition to preserving the visual identity of the speaker. Modeling this audio-visual correspondence remains a fundamental challenge because speech provides only implicit cues about facial motion, while its visual realization varies with speaker-specific facial geometry and articulation patterns.

Existing methods generally follow either an end-to-end or a multi-stage generation paradigm. End-to-end methods Chen et al. (2025); Ji et al. (2025) directly synthesize talking head videos from the driving audio and a reference image or video. However, the correspondence between speech cues and facial dynamics is modeled implicitly within the video generator, making fine-grained audio-visual alignment difficult to explicitly constrain. To achieve more structured motion control, multi-stage methods Zhong et al. (2023); Wei et al. (2024); Kim et al. (2025) employ intermediate motion representations, such as facial landmarks, to bridge speech conditioning and video synthesis. However, such intermediate motion representations cannot fully retain fine-grained articulation cues, making it difficult for subsequent synthesis to faithfully reproduce speech-related visual correspondence.

This limitation gives rise to two key challenges. First, speech-driven facial motion is identity dependent, yet existing audio-to-motion models typically incorporate identity through feature modulation Wei et al. (2024); Tan et al. (2024) or attention Kim et al. (2025); Li et al. (2024) within a geometric predictor. In such formulations, identity mainly serves as a conditioning signal for facial geometry, rather than jointly participating with speech in modeling the temporal evolution of fa-

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.

cial motion, which limits the capture of speaker-specific articulation patterns. Second, audio-visual correspondence should be adapted as the representation progresses from speech to landmarks and ultimately to rendered video. However, existing pipelines typically propagate speech-related motion cues mainly through the previously predicted landmarks, making subsequent synthesis largely dependent on the correspondence already captured during initial motion prediction. Therefore, fine-grained errors in speech-motion correspondence can propagate to the final video, causing the rendered facial dynamics to deviate from the driving speech. As a result, overcoming the above limitations requires the joint modeling of speech and speaker-specific visual cues, together with persistent audio guidance throughout motion prediction and video synthesis.

To address these challenges, we propose **PlanTalk**, an identity-aware cross-modal motion planning framework with hierarchical audio guidance for talking head generation. **First**, to move beyond treating speaker identity merely as an auxiliary condition for geometric prediction, we reformulate facial motion generation as an identity-aware cross-modal planning problem rather than direct landmark regression. This formulation requires the joint interpretation of speech and identity cues, together with sequence-level organization of facial dynamics. Multimodal Large Language Models(MLLMs) are particularly well suited to this formulation because they unify multimodal contextual modeling and autoregressive sequence generation within a common architecture. Accordingly, PlanTalk introduces an MLLM-based motion planner that jointly interprets the driving speech and speaker-specific visual characteristics and autoregressively organizes them into an identity-aware facial motion sequence. In this way, identity does not simply modulate the predicted geometry; instead, the joint speech-identity context directly participates in determining how facial articulation evolves over time. **Second**, to prevent audio-visual correspondence from being merely inherited from initial motion prediction, PlanTalk introduces hierarchical audio guidance across subsequent representation levels. At the landmark level, an audio-landmark fusion module further reintegrates fine-grained acoustic cues into the planned motion, refining the temporal and geometric correspondence between facial articulation and speech. During video synthesis, denoising-stage audio attention directly couples the driving audio with the visual representation, enabling the diffusion process to preserve speech-dependent facial dynamics that cannot be fully specified by landmarks alone. Our principle contributions can be summarized as follows:

- We formulate audio-driven facial motion generation as an **identity-aware cross-modal motion planning** problem, moving beyond conventional audio-to-motion regression based on feature-level conditioning. Based on this formulation, we propose PlanTalk, which leverages an MLLM to jointly model speech context and speaker-specific visual characteristics for identity-aware facial motion generation.
- We introduce a **hierarchical audio-guidance strategy** that progressively refines and preserves audio-visual correspondence across the generation pipeline. Specifically, audio-landmark fusion improves fine-grained synchronization at the motion level, while denoising-stage audio attention maintains direct speech guidance during video synthesis.
- Extensive experiments on talking head datasets demonstrate that PlanTalk achieves superior lip synchronization and identity consistency while maintaining high visual fidelity. Further analyses validate the effectiveness of MLLM-based cross-modal motion planning and each level of audio guidance.

2 Related Work

Audio-driven talking head generation aims to synthesize realistic facial videos whose motion is synchronized with driving speech. Existing methods can be broadly categorized according to how audio information is translated into visual dynamics: end-to-end audio-to-video generation and multi-stage generation through intermediate motion representations.

End-to-End Talking Head Generation. Early methods commonly condition facial animation on acoustic features to synthesize facial frames that are visually consistent with the driving speech Zhou et al. (2021); Wang et al. (2021a); Ji et al. (2021). With the development of diffusion models and other powerful generative architectures Rombach et al. (2022); Peebles & Xie (2023), recent approaches have substantially improved visual quality and temporal coherence. EchoMimic Chen et al. (2025), Sonic Ji et al. (2025), and Hallo2 Cui et al. (2025), for example, apply diffusion-based video generators with driving audio and reference appearance to synthesize complete talking head

sequences. These methods benefit from jointly optimizing motion generation and visual rendering within a unified framework. However, the relationship between speech cues and facial dynamics is represented implicitly in the high-dimensional generation process. Consequently, they provide limited explicit control over how phonetic and temporal information should be expressed through fine-grained facial movements, making subtle audio-visual misalignment difficult to identify and correct.

Multi-Stage Talking Head Generation. To introduce explicit and interpretable motion control, multi-stage methods Gan et al. (2023); Li et al. (2025); Wei et al. (2024); Kim et al. (2025) decompose talking head generation into motion prediction and appearance synthesis. The intermediate representation take the form of facial landmarks, 3D facial coefficients, or other structured motion descriptors. Among these alternatives, facial landmarks provide a compact representation of facial geometry and can effectively separate motion dynamics from appearance information. AniPortrait Wei et al. (2024) derives 2D landmarks from audio through 3D facial representations and uses the resulting trajectories to guide diffusion-based video generation. MoDiTalker Kim et al. (2025) predicts lip-synchronized facial landmarks using an Audio-to-Motion module and subsequently employs them as geometric conditions for video synthesis. By externalizing facial motion from the video generator, these methods offer stronger structural control than end-to-end generation methods. However, despite providing explicit intermediate control, existing multi-stage methods largely rely on geometrically predicted landmarks, which offer relatively coarse motion guidance for bridging speech conditioning and video synthesis. In contrast, PlanTalk retains the geometric controllability of multi-stage generation while incorporating MLLM-based identity-aware motion modeling and continued speech guidance throughout subsequent stages.

3 METHODOLOGY

As illustrated in Fig. 1, PlanTalk consists of two designs: *identity-aware cross-modal motion planning* and *hierarchical audio guidance*. The former employs an MLLM to jointly interpret the driving audio and speaker-specific visual characteristics, organizing them into an identity-aware facial motion sequence. The latter progressively adapts audio-visual correspondence to different representation levels. Specifically, audio-landmark fusion further integrates acoustic and landmark representations to improve the fine-grained alignment between facial motion and speech articulation, while denoising-stage audio attention directly conditions video synthesis on the driving speech.

3.1 IDENTITY-AWARE CROSS-MODAL MOTION PLANNING

Speech-driven facial motion depends not only on acoustic content but also on speaker-specific facial geometry and articulation patterns. Instead of treating identity as an auxiliary condition for landmark regression, we formulate facial motion generation as a cross-modal sequence-planning problem, in which speech and identity jointly determine the temporal evolution of facial articulation.

3.1.1 MULTIMODAL REPRESENTATION

Given the identity frame f_I and driving audio A, we employ the visual and audio encoder to extract their modality-specific representations: $\mathbf{E}_I = \mathcal{E}_I(f_I)$ and $\mathbf{E}_A = \mathcal{E}_A(A)$, where $\mathcal{E}_I$ and $\mathcal{E}_A$ denote the image and the audio encoder, respectively. To align the modality-specific features with the token space of the language model, we introduce trainable visual and audio projection layers: $\mathbf{T}_I = \mathcal{P}_I(\mathbf{E}_I)$ and $\mathbf{T}_A = \mathcal{P}_A(\mathbf{E}_A)$. The resulting representations encode visual identity and temporal speech context in a form that can be jointly processed by the MLLM.

3.1.2 AUTOREGRESSIVE MOTION PLANNING

Conditioned on $\mathbf{T}_I$ and $\mathbf{T}_A$, the MLLM autoregressively produces a sequence of landmark hidden states $\mathbf{H}_L = \{t_L^n\}_{n=1}^N$:

$$p(\mathbf{H}_L \mid f_I, A) = \prod_{n=1}^{N} p\left(t_L^n \mid t_L^{<n}, \mathbf{T}_I, \mathbf{T}_A\right). \tag{1}$$

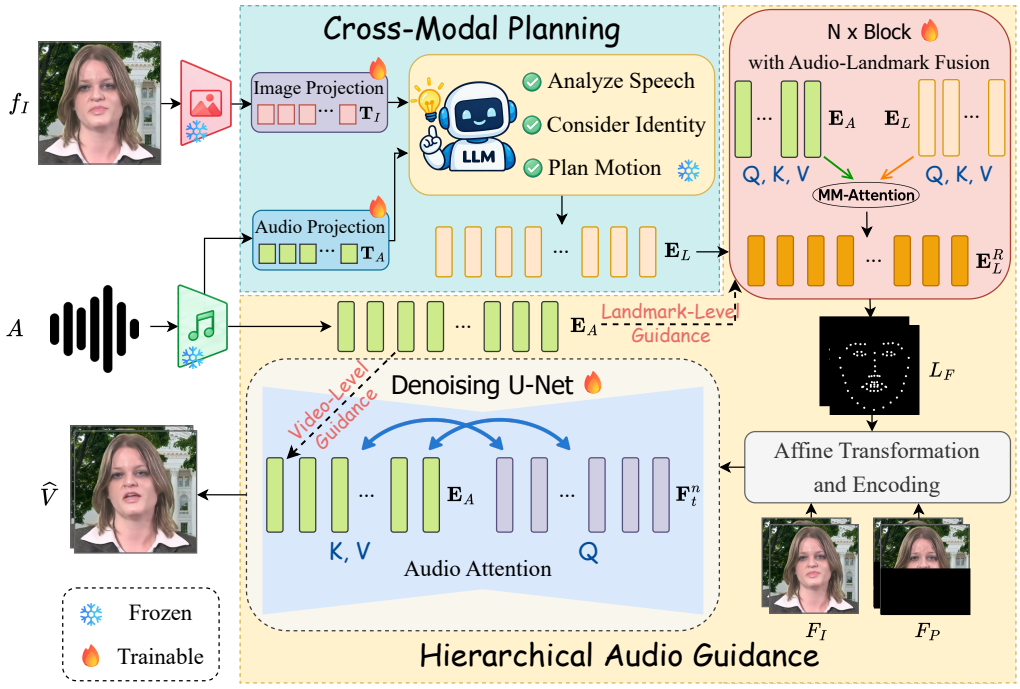

Figure 1: **Overall network architecture of PlanTalk.** Given an identity frame f_I and driving audio A, the MLLM-based planner generates an identity-aware landmark representation. Hierarchical audio guidance subsequently refines audio-visual correspondence at two levels: audio-landmark fusion produces speech-aligned landmarks L_F through multimodal joint attention, while audio attention directly injects speech information into the denoising UNet. Together with identity frames F_I and pose frames F_P, the refined landmarks guide the synthesis of the final video $\widehat{V}$.

Unlike frame-wise geometric regression, Eq. (1) conditions each motion token on the joint speech-identity context and previously generated motion tokens. The resulting sequence therefore captures both speaker-specific articulation and temporal motion dependencies. The output hidden states are mapped into landmark embeddings through a trainable projection layer $\mathbf{E}_L = \mathcal{P}_L(\mathbf{H}_L)$, where $\mathbf{E}_L$ represents the planned facial motion embeddings and serves as the input to hierarchical audio guidance. In this way, the MLLM establishes a sequence-level motion plan for subsequent refinement.

3.2 Hierarchical Audio Guidance

Cross-modal planning models the global relationship among speech, identity, and facial motion. However, the overall planned landmark representation cannot fully reflect acoustic variations associated with subtle articulation. Moreover, landmarks alone cannot completely specify high-dimensional speech-related appearance during video synthesis. We therefore introduce hierarchical audio guidance across two successive representation levels: landmark refinement and video synthesis.

3.2.1 Audio-Landmark Fusion

We employ multimodal joint attention(MM-Attention) Cheng et al. (2025); Esser et al. (2024); Labs (2024) to model bidirectional interactions between acoustic and geometric features. Given the landmark embeddings $\mathbf{E}_L$ and the audio embeddings $\mathbf{E}_A$, we first compute modality-specific query, key, and value representations:

$$\mathbf{Q}_m = \mathbf{E}_m \mathbf{W}_{\mathbf{Q}}^m, \qquad \mathbf{K}_m = \mathbf{E}_m \mathbf{W}_{\mathbf{K}}^m, \qquad \mathbf{V}_m = \mathbf{E}_m \mathbf{W}_{\mathbf{V}}^m, \qquad m \in \{L, A\}. \tag{2}$$

We then concatenate the corresponding representations and perform joint self-attention:

$$\mathbf{Q} = [\mathbf{Q}_L; \mathbf{Q}_A], \qquad \mathbf{K} = [\mathbf{K}_L; \mathbf{K}_A], \qquad \mathbf{V} = [\mathbf{V}_L; \mathbf{V}_A], \qquad \mathbf{E} = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}. \tag{3}$$

The landmark portion in $\mathbf{E}$ is then linearly projected and combined with a residual connection to obtain the refined landmark embedding $\mathbf{E}_L^R$, followed by the landmark decoder to generate the refined landmark L_F. Since MM-Attention allows the network to dynamically weight the importance of audio and landmark features, the refined representation incorporates acoustic context relevant to the corresponding articulation state, yielding finer-grained audio-motion correspondence.

3.2.2 AUDIO-CONDITIONED VIDEO SYNTHESIS

We construct the video generator based on the Motion-to-Video architecture of MoDiTalker Kim et al. (2025). The refined landmarks L_F are first processed by an affine transformation and then encoded together with F_I and F_P as motion, identity, and pose conditions for video synthesis. Unlike existing landmark-driven synthesis pipelines, PlanTalk retains direct access to the driving audio throughout denoising. To provide temporally continuous speech guidance, the audio representation associated with frame n is augmented using its local neighborhood $\overline{\mathbf{E}}_A^n = \left[\mathbf{E}_A^{n-r}; \ldots; \mathbf{E}_A^n; \ldots; \mathbf{E}_A^{n+r}\right]_{\text{seq}}$, where r determines the temporal context range. Let $\mathbf{F}_t^n$ denote the UNet intermediate visual feature for frame n at diffusion timestep t. We introduce cross-attention between the visual latent and the local audio context:

$$\mathbf{Q}_t^n = \mathbf{F}_t^n \mathbf{W}_{\mathbf{Q}}, \qquad \mathbf{K}_A^n = \overline{\mathbf{E}}_A^n \mathbf{W}_{\mathbf{K}}, \qquad \mathbf{V}_A^n = \overline{\mathbf{E}}_A^n \mathbf{W}_{\mathbf{V}}, \tag{4}$$

$$\widehat{\mathbf{F}}_t^n = \mathbf{F}_t^n + \text{Softmax}\left(\frac{\mathbf{Q}_t^n (\mathbf{K}_A^n)^\top}{\sqrt{d}}\right) \mathbf{W}_O \mathbf{V}_A^n. \tag{5}$$

This audio-attention mechanism allows the evolving video representation to directly access speech cues during denoising. It therefore complements landmark-based geometric control with acoustic information, providing additional guidance for speech-dependent visual details during video synthesis and ultimately producing the final video $\widehat{V}$.

3.3 TRAINING OBJECTIVES

PlanTalk is supervised at both the landmark and video levels. Landmark supervision trains the MLLM planner and the audio-landmark fusion module to generate identity-consistent and temporally coherent motion. Video-level supervision optimizes the condition encoders for video synthesis and the audio-conditioned diffusion model.

3.3.1 LANDMARK OBJECTIVE

Let $\mathbf{L} = \{\mathbf{L}_n\}_{n=1}^N$ and $\widehat{\mathbf{L}} = \{\widehat{\mathbf{L}}_n\}_{n=1}^N$ denote the ground-truth and predicted landmark sequences, respectively. We use the first-frame landmark $\mathbf{L}_1$ as the identity-related geometric reference and define the motion residuals as $\mathbf{R}_n = \mathbf{L}_n - \mathbf{L}_1$, and $\widehat{\mathbf{R}}_n = \widehat{\mathbf{L}}_n - \widehat{\mathbf{L}}_1$. The landmark reconstruction and motion-residual losses are $\mathcal{L}_{\text{final}} = \rho(\widehat{\mathbf{L}}, \mathbf{L})$ and $\mathcal{L}_{\text{res}} = \rho(\widehat{\mathbf{R}}, \mathbf{R})$, respectively, where $\rho(\cdot, \cdot)$ denotes the robust ℓ_1 loss. To supervise facial dynamics across multiple temporal ranges, we define the k-frame temporal difference as $\Delta_k \mathbf{R}_n = \mathbf{R}_{n+k} - \mathbf{R}_n$, and formulate the multi-scale velocity loss as:

$$\mathcal{L}_{\text{vel}} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \rho\left(\Delta_k \widehat{\mathbf{R}}, \Delta_k \mathbf{R}\right), \qquad \mathcal{K} = \{1, 2, 4\}. \tag{6}$$

These temporal intervals supervise short-, medium-, and relatively long-term motion dynamics. We further constrain the initial facial geometry using $\mathcal{L}_{\text{id}} = \rho\left(\widehat{\mathbf{L}}_1, \mathbf{L}_1\right)$. The complete landmark objective is:

$$\mathcal{L}_{\text{landmark}} = \mathcal{L}_{\text{final}} + \mathcal{L}_{\text{res}} + 0.5\mathcal{L}_{\text{vel}} + 0.5\mathcal{L}_{\text{id}}. \tag{7}$$

3.3.2 ENCODING OBJECTIVE

The landmark, identity, and pose conditions are encoded by $\mathcal{E}_L$, $\mathcal{E}_I$, and $\mathcal{E}_P$, respectively. Following the encoding strategy of the underlying Motion-to-Video model Kim et al. (2025), these encoders are optimized using reconstruction and perceptual supervision:

$$\mathcal{L}_{\text{encoder}} = \mathcal{L}_{\text{reconstruct}} + \mathcal{L}_{\text{percept}}, \tag{8}$$

where $\mathcal{L}_{\text{reconstruct}}$ is the pixel-level reconstruction loss Zhao et al. (2017), and $\mathcal{L}_{\text{percept}}$ is the perceptual loss Zhang et al. (2018).

3.3.3 AUDIO-CONDITIONED DIFFUSION OBJECTIVE

The refined landmarks, identity frames, and target pose frames are encoded as $\mathbf{E}'_L = \mathcal{E}_L(L_F)$, $\mathbf{E}_I = \mathcal{E}_I(F_I)$, and $\mathbf{E}_P = \mathcal{E}_P(F_P)$. Together with audio embeddings $\mathbf{E}_A$, they constitute the diffusion conditions. Following the diffusion formulation, the denoising network reconstructs the clean video latent using:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{V,t,\boldsymbol{\epsilon},\mathbf{E}'_L,\mathbf{E}_I,\mathbf{E}_P,\mathbf{E}_A}\left[\left\|V_0 - \mathcal{F}_{\text{diffusion}}\left(V_t, t; \mathbf{E}'_L, \mathbf{E}_I, \mathbf{E}_P, \mathbf{E}_A\right)\right\|_2^2\right], \tag{9}$$

where V_0 is the ground-truth talking head video, and V_t is the noised V_0 at time step t. The overall training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{landmark}} + \mathcal{L}_{\text{encoder}} + \mathcal{L}_{\text{diffusion}}. \tag{10}$$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Implementation details. We apply the pretrained ImageBind Girdhar et al. (2023) as the unified image and audio encoder in MLLM, and take the Vicuna-7B Chiang et al. (2023) as the Large Language Model(LLM) backbone, following its widespread adoption in existing MLLMs Su et al. (2023); Zhang et al. (2023a). During training, the Vicuna-7B and the ImageBind encoder are kept frozen, while the remaining trainable modules are jointly optimized. The input image size for MLLM is 256×256. The model is trained using the Adam optimizer with a learning rate of 1e-4 and a weight decay of 0.01. The gradient clipping is applied with a maximum gradient norm of 0.5, and we train with a per-device batch size of 2. All experiments are conducted on 4 NVIDIA A100 80GB GPUs.

Datasets. We train our model on the LRS3-TED Afouras et al. (2018), HDTF Zhang et al. (2021), and CelebV-HQ Zhu et al. (2022) datasets. LRS3-TED provides a large-scale lip-reading corpus comprising approximately 400 hours of TED videos. HDTF contains 16 hours of high-resolution videos featuring more than 300 subjects, while CelebV-HQ consists of 35,666 video clips covering 15,653 identities. For evaluation, we randomly select 98 videos from HDTF and 50 videos from CelebV-HQ, using the remaining videos for training. The facial landmarks required during training are extracted using a 3D Morphable Model (3DMM) Blanz & Vetter (2023).

Evaluation Metrics. We evaluate the generated videos using five widely adopted metrics. Specifically, **FID** Heusel et al. (2017) and **PSNR** Hore & Ziou (2010) assess visual quality and reconstruction fidelity, respectively, while **CSIM** Deng et al. (2019) measures identity preservation. To evaluate audio-visual synchronization, we report **LSE-D** Chung & Zisserman (2016) which quantifies the distance between audio and lip-motion features. We further employ **LMD** Chen et al. (2018) to measure the accuracy of the generated lip movements.

Baselines. We compare PlanTalk with representative talking-head generation methods from two main paradigms: end-to-end approaches, including Sonic Ji et al. (2025) and EchoMimic Chen et al. (2025), and multi-stage approaches, including SadTalker Zhang et al. (2023b), AniPortrait Wei et al. (2024), SyncTalk Peng et al. (2024), and MoDiTalker Kim et al. (2025).

4.2 QUANTITATIVE RESULTS

Results on HDTF and CelebV-HQ testsets. Table 1 presents the quantitative comparisons of PlanTalk and the competitors on both HDTF and CelebV-HQ datasets. PlanTalk achieves competitive performance across all evaluation metrics. The most pronounced advantages of PlanTalk are observed in LSE-D and LMD. The leading performance on both metrics indicates that PlanTalk can more accurately capture fine-grained speech-related facial dynamics, compared with strong baselines such as Sonic and MoDiTalker, which are specifically designed for audio-motion alignment.

In terms of video quality, PlanTalk also demonstrates great performance, achieving the highest PSNR and CSIM scores while remaining highly competitive in FID. These results further show that the improved motion accuracy is not achieved at the expense of visual quality or identity consistency, suggesting that PlanTalk can effectively separate dynamic facial attributes from relatively static appearance information. Its overall performance reflects a more meaningful balance: the generated

Table 1: **Quantitative comparison results on both the HDTF and CelebV-HQ datasets.** We evaluate video quality and lip-synchronization accuracy on 98 HDTF videos and 50 CelebV-HQ videos. The best and second-best performances are marked in red and orange, respectively.

	Video Quality			Lip Sync.	
HDTF Testset	FID ↓	PSNR ↑	CSIM ↑	LMD ↓	LSE-D ↓
SadTalker Zhang et al. (2023b)	25.48	29.83	0.81	2.22	9.96
AniPortrait Wei et al. (2024)	37.07	33.86	0.75	1.92	10.36
SyncTalk Peng et al. (2024)	26.69	30.95	0.66	3.58	10.57
EchoMimic Chen et al. (2025)	22.41	34.68	0.79	2.01	10.89
Sonic Ji et al. (2025)	20.45	34.73	0.80	2.07	9.67
MoDiTalker Kim et al. (2025)	14.15	35.82	0.92	1.38	9.15
PlanTalk(Ours)	14.38	36.21	0.97	1.19	9.02
CelebV-HQ Testset					
SadTalker Zhang et al. (2023b)	27.52	28.77	0.79	5.38	11.12
AniPortrait Wei et al. (2024)	36.35	33.52	0.76	6.73	11.80
SyncTalk Peng et al. (2024)	26.18	34.32	0.76	5.48	11.00
EchoMimic Chen et al. (2025)	28.59	34.65	0.72	5.79	10.77
Sonic Ji et al. (2025)	23.64	35.77	0.78	5.39	10.57
MoDiTalker Kim et al. (2025)	18.93	36.39	0.83	4.94	10.63
PlanTalk(Ours)	19.00	37.03	0.85	4.79	9.94

Table 2: Efficiency Comparison (Inference time for generating a video with 5.76 seconds). The best and second-best performances are marked in red and orange, respectively.

SadTalker	AniPortrait	SyncTalk	Echomimic	Sonic	MoDiTalker	**PlanTalk(Ours)**
46.1s	428.7s	22.0s	481.3s	38.0s	23.7s	20.1s

videos are not only visually realistic but also exhibit strong temporal synchronization, accurate lip geometry, and consistent speaker identity.

Computational complexity. As shown in Table 2, PlanTalk provides higher generation efficiency than the compared diffusion-based talking head methods. This result is noteworthy because the introduction of an MLLM-based planning module might intuitively be expected to increase inference cost. However, the computational overhead of PlanTalk is governed not only by the presence of an additional reasoning component, but also by how frequently that component is executed. This difference becomes particularly clear when compared with MoDiTalker, a multi-stage method with competitive inference efficiency that likewise uses predicted landmarks as structured guidance, but employs a diffusion-based motion generator to synthesize landmark trajectories through iterative denoising. PlanTalk instead directly generates landmark representations from multimodal inputs, avoiding this denoising overhead associated with motion diffusion, while the resulting landmarks can be directly used as structured guidance for subsequent stages. Consequently, introducing MLLM-based motion planning does not incur the iterative sampling cost of diffusion-based motion generation, allowing PlanTalk to maintain competitive overall inference efficiency despite the additional cross-modal reasoning module.

4.3 Ablation Study

To assess the contribution of each key component, we conduct comprehensive ablation studies on the HDTF testset. Specifically, we evaluate the effects of audio-landmark fusion and audio attention by removing each component individually. The qualitative results are presented in Table 3.

Table 3: Ablation Study on PlanTalk with the HDTF testset.

Fusion	Audio Attention	CSIM ↑	LMD ↓	LSE-D ↓
×	×	0.89	1.28	9.17
×	✓	0.91	1.24	9.10
✓	×	0.96	1.21	9.07
✓	✓	**0.97**	**1.19**	**9.02**



Figure 2: Qualitative comparison of frames generated by our method and the baseline methods.

Effect of Audio Attention. When the audio-landmark fusion blocks are not applied, introducing the audio attention into the denoising UNet leads to consistent but relatively modest improvements, as shown in Table 3. In particular, LMD and LSE-D show a noticeable reduction because they are more sensitive to audio synchronization, while CSIM exhibits a slight yet stable gain, indicating that explicitly incorporating audio representations into video denoising mainly enhances lip synchronization by providing complementary speech guidance beyond the landmark condition. This further confirms the importance of maintaining audio constraints during video synthesis rather than terminating them at the intermediate motion representation.

Effect of Audio-Landmark Fusion. Table 3 shows that when the audio attention is removed, the audio-landmark fusion module alone significantly improves all evaluation metrics, especially in CSIM. This demonstrates that the audio-landmark fusion mechanism, which further models the interaction between acoustic cues and facial geometry, can enable the predicted lip movements to better follow speech articulation while effectively preserving identity consistency.

When both the audio attention and the audio-landmark fusion module are enabled, the model achieves the best performance. This result reflects the progressive gains across different representation levels: audio-landmark fusion first improves articulation accuracy while retaining identity-related facial characteristics at the landmark level, and audio attention subsequently reinforces speech consistency during video synthesis, yielding further overall improvements. Together, they

validate the effectiveness of the hierarchical audio-guidance design in PlanTalk across the generation pipeline.

4.4 Qualitative Results

Qualitative results are shown in Fig. 2. Across different identities and speech segments, PlanTalk generates facial motions that more closely follow the ground-truth articulation while maintaining the stable visual appearance. In terms of lip synchronization, our method produces more accurate mouth shapes and more consistent lip-motion transitions across frames. In contrast, competing methods (e.g., EchoMimic, MoDiTalker) exhibit noticeable mismatches in mouth opening and lip configurations at several timestamps, indicating less precise audio-motion correspondence. Regarding visual quality and identity preservation, PlanTalk produces more faithful facial details, particularly around the mouth and teeth regions, while maintaining consistent speaker identity. By comparison, other methods (e.g., SyncTalk, Sonic) show local blurring and occasional identity inconsistencies. These qualitative observations are consistent with the quantitative improvements in lip synchronization and identity preservation.

To investigate the importance of cross-modal reasoning of MLLM for motion planning, we compare the facial landmark trajectories generated by the MLLM directly with those predicted by the Audio-to-Motion model in MoDiTalker Kim et al. (2025). As shown in Fig. 3, the landmarks produced by the MLLM exhibit more evident and temporally coherent variations across frames, particularly around the mouth region, and better follow the corresponding facial dynamics. In contrast, the MoDiTalker model produces relatively limited motion variations and captures fewer fine-grained changes in speech-related facial geometry. These observations suggest that MLLM-based cross-modal planning provides a more expressive mechanism for modeling the interaction between speech and speaker-specific visual cues than conventional audio-conditioned motion prediction, resulting in more plausible and temporally coherent facial motion trajectories.

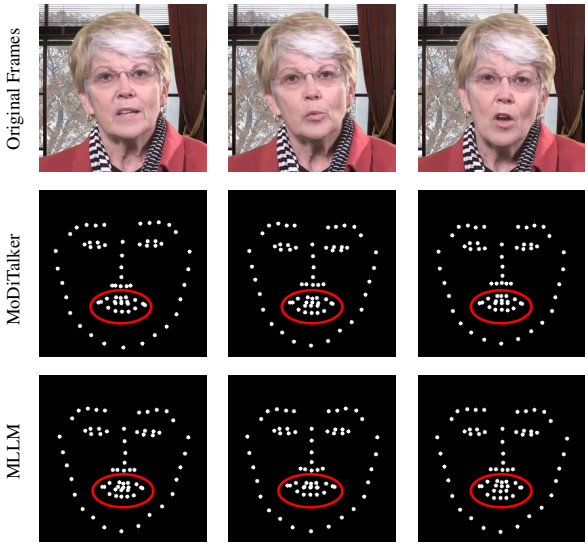

Figure 3: Qualitative comparison of facial landmark trajectories (in face-centric coordinates) generated by MLLM directly and the Audio-to-Motion model in MoDiTalker.

5 Conclusion

In this paper, we present PlanTalk, a unified framework for audio-driven talking head generation that integrates MLLM-based identity-aware cross-modal motion planning with hierarchical audio guidance. Rather than relying on conventional speech-conditioned geometric prediction, PlanTalk formulates facial motion generation as a cross-modal planning problem and maintains speech guidance throughout the whole generation pipeline, progressively establishing, refining, and preserving audio-visual correspondence from motion planning to video synthesis. This formulation enables the model to jointly exploit speech and speaker-specific visual information for fine-grained and speech-synchronized facial dynamics while maintaining identity consistency. Extensive experiments demonstrate that PlanTalk achieves state-of-the-art performance in lip synchronization and identity consistency while maintaining high visual fidelity.

REFERENCES

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 157–164. 2023.

Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *European Conference on Computer Vision*, pp. 538–553, 2018.

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

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28901–28911, 2025.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian Conference on Computer Vision*, pp. 251–263, 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. In *International Conference on Learning Representations*, pp. 91659–91671, 2025.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22577–22588, 2023.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Tianyu He, Junliang Guo, Runyi Yu, Yuchi Wang, Kaikai An, Leyi Li, Xu Tan, Chunyu Wang, Han Hu, HsiangTao Wu, et al. Gaia: Zero-shot talking avatar generation. In *International Conference on Learning Representations*, pp. 2107–2124, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *International Conference on Pattern Recognition*, pp. 2366–2369, 2010.

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

Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14075–14084, 2021.

Seyeon Kim, Siyoon Jin, Jihye Park, Kihong Kim, Jiyoung Kim, Jisu Nam, and Seungryong Kim. Moditalker: Motion-disentangled diffusion model for high-fidelity talking head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4302–4310, 2025.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Dongze Li, Kang Zhao, Wei Wang, Yifeng Ma, Bo Peng, Yingya Zhang, and Jing Dong. S 3 d-nerf: single-shot speech-driven neural radiance field for high fidelity talking head synthesis. In *European Conference on Computer Vision*, pp. 365–382, 2024.

Tianqi Li, Ruobing Zheng, Minghui Yang, Jingdong Chen, and Ming Yang. Ditto: Motion-space diffusion for controllable realtime talking head synthesis. In *Proceedings of the ACM International Conference on Multimedia*, pp. 9704–9713, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4172–4182, 2023.

Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 666–676, 2024.

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 ACM International Conference on Multimedia*, pp. 484–492, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 10674–10685, 2022.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the Era of Interactive Assistants!*, pp. 11–23, 2023.

Shuai Tan, Bin Ji, Yu Ding, and Ye Pan. Say anything with any style. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5088–5096, 2024.

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

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10034–10044, 2021b.

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

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, 2023a.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

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*, pp. 8652–8661, 2023b.

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

Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2017.

Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2023.

Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4174–4184, 2021.

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 Appendix

In Section A.1, we provide a detailed description of the complete inference pipeline of PlanTalk, illustrating how the proposed components are integrated from cross-modal motion planning to final video synthesis. In Section A.2, we provide additional details on the training of the landmark, identity, and pose encoders following the audio-landmark fusion stage. In Section A.3, we further present additional qualitative comparisons with existing methods to provide a more comprehensive visual assessment of PlanTalk and further demonstrate its effectiveness in generating audio-synchronized talking-head videos with high quality. In Section A.4, we clarify the AI assistance in the paper. Additionally, we provide several demo videos in the supplementary material.

A.1 Inference

During inference, the first identity frame f_I and driving audio A are processed by the MLLM-based planner to obtain an identity-aware motion representation. Audio-landmark fusion subsequently refines the planned motion into a speech-aligned landmark sequence. After pose transformation, the landmarks are encoded together with the identity and pose conditions to guide diffusion-based video synthesis. The audio representation remains available throughout landmark refinement and denoising, thereby maintaining speech guidance across the entire generation pipeline. The detailed inference pipeline is summarized in Algorithm 1.

Algorithm 1 PlanTalk Inference Pipeline

Input: Identity frames F_I, Audio A, Pose frames F_P
Run: PlanTalk for talking head video generation
Extract the first identity frame and the corresponding landmark: $f_I \leftarrow F_I$
Identity features through encoder: $\mathbf{E}_I \leftarrow \text{Encoder}(f_I)$
Audio features through encoder: $\mathbf{E}_A \leftarrow \text{Encoder}(A)$
Identity token embeddings: $\mathbf{T}_I \leftarrow \text{Project}(\mathbf{E}_I)$
Audio token embeddings: $\mathbf{T}_A \leftarrow \text{Project}(\mathbf{E}_A)$
Output landmark embeddings: $\mathbf{E}_L \leftarrow \text{Project}(\text{LLM}(\{\mathbf{T}_I, \mathbf{T}_A\}))$
Refine landmark embeddings with audio embeddings: $\mathbf{E}_L^R \leftarrow \text{Fusion}(\mathbf{E}_L, \mathbf{E}_A)$
Output landmarks after fusion: $L_F \leftarrow \text{Decoder}(\mathbf{E}_L^R)$
Obtain synthesized landmarks through the affine transformation: $L \leftarrow \text{Affine}(L_F, F_P)$
Concat landmark, pose, and identity embeddings: $\mathbf{E}_C \leftarrow \text{Concat}(\text{Encoder}(L, F_P, F_I))$
Generate the video through diffusion model: $\widehat{V} \leftarrow \text{Diffusion}(\mathbf{E}_C, \mathbf{E}_A)$
Output: Sequence of generated frames $\widehat{V}$

A.2 More Details of Training Encoders

After the audio-landmark fusion stage, PlanTalk employs three condition encoders for subsequent video synthesis: the landmark encoder $\mathcal{E}_L$, identity encoder $\mathcal{E}_I$, and pose encoder $\mathcal{E}_P$, which encode the landmark, identity, and pose conditions, respectively. For notational simplicity, we use G as a generic input variable to uniformly describe the training of these three encoders, rather than referring to a specific video sequence. Specifically, G denotes the spatial landmark representation when training $\mathcal{E}_L$, the identity frames F_I when training $\mathcal{E}_I$, and the pose frames F_P when training $\mathcal{E}_P$. Each input is represented as a cubic-like 4D tensor $G \in \mathbb{R}^{S\times C\times H\times W}$, where S, C, H, and W denote the sequence length, number of channels, height, and width, respectively. The corresponding encoder compresses G into a tri-plane latent representation consisting of three image-like 2D feature planes, each of size $\mathbb{R}^{c\times h\times w}$, where c, h, and w denote the latent channel dimension, height, and width.

To simplify the following formulation, we use $\mathcal{E}$ to denote any of the three encoders $\{\mathcal{E}_L, \mathcal{E}_I, \mathcal{E}_P\}$ and $\mathcal{D}$ to denote its corresponding decoder. Each encoder-decoder pair is trained in an autoencoding manner: $\mathcal{E}$ maps the corresponding input G from the RGB space into the latent space, while $\mathcal{D}$ reconstructs the latent representation back to $\widehat{G}$ in the RGB space. During training, we jointly employ a pixel-level reconstruction loss Zhao et al. (2017) and a perceptual loss Zhang et al. (2018) to encourage faithful reconstruction in both the pixel and perceptual feature spaces. The overall training objective is formulated as:

$$\mathcal{L}_{\text{encoder}} = \mathbb{E}_G\left[\left\|G-\widehat{G}\right\|_1\right] + \mathbb{E}_G\left[\left\|\phi(G)-\phi(\widehat{G})\right\|_1\right], \tag{11}$$

where $\phi(\cdot)$ denotes the feature extractor used to compute the perceptual loss.

A.3 More Qualitative Comparisons

In this subsection, we provide additional visual comparisons of our PlanTalk with the recent competitors, including end-to-end approaches: Sonic Ji et al. (2025) and EchoMimic Chen et al. (2025), and multi-stage approaches: SadTalker Zhang et al. (2023b), AniPortrait Wei et al. (2024), SyncTalk Peng et al. (2024), and MoDiTalker Kim et al. (2025). The comparison results are shown in Fig. 4.

A.4 AI Assistance Statement

LLM-based tools are used only for language polishing and improving the clarity of the manuscript. All technical content, experiments, analyses, and conclusions were developed and verified by the authors.



Figure 4: More qualitative comparison of generated frames with different baseline methods. It can be observed that our method significantly outperforms the competitors in both lip synchronization and identity consistency.