JAM-Flow: Joint Audio-Motion Synthesis with Flow Matching

Mingi Kwon[1,2⋆], Joonghyuk Shin[3⋆], Jaeseok Jeong[1,2], Jaesik Park[3†], and Youngjung Uh[1†]

[1] Yonsei University, Seoul, Republic of Korea
[2] TheFamiliarLab, Toronto, Canada
[3] Seoul National University, Seoul, Republic of Korea
{kwonmingi,jete_jeong,yj.uh}@yonsei.ac.kr
{joonghyuk,jaesik.park}@snu.ac.kr

Abstract. The intrinsic link between facial motion and speech is often overlooked in generative modeling, where talking head synthesis and text-to-speech (TTS) are typically addressed as separate tasks. This paper introduces JAM-Flow, a unified framework to simultaneously synthesize and condition on both facial motion and speech. Our approach leverages flow matching and a novel Multi-Modal Diffusion Transformer architecture, integrating specialized Motion-DiT and Audio-DiT modules. These are coupled via selective joint attention layers and incorporate key architectural choices, such as temporally aligned positional embeddings and localized joint attention masking, to enable effective cross-modal interaction while preserving modality-specific strengths. By analyzing and leveraging pretrained representation embeddings, JAM-Flow is designed for efficient, near real-time sampling. Trained with an inpainting-style objective, JAM-Flow supports a wide array of conditioning inputs (including text, reference audio, and reference motion) facilitating tasks such as synchronized talking head generation from text, audio-driven animation, and much more, within a single, coherent model. JAM-Flow significantly advances multi-modal generative modeling by providing a practical solution for holistic audio-visual synthesis. Project website

Keywords: Joint Audio-Motion Generation · Talking Head Generation · Text-to-Speech

1 Introduction

With the rapid advancement of generative models [15, 23, 33, 34, 44, 50], the synthesis of realistic human faces and voices has become increasingly sophisticated. Two major fields have emerged from this trend: talking head generation [9, 19, 32, 49, 53, 57, 63, 67, 69, 80], which animates static portrait images to mimic facial expressions, and text-to-speech (TTS) synthesis [3, 10, 11, 18, 26], which converts

⋆Equal contribution. † Corresponding Authors.

text and a short voice reference into natural-sounding speech. While state-of-the-art methods in both domains, ranging from GAN-based models [19, 77] for near real-time inference to diffusion-based models [32, 67] and flow matching-based models [3, 26] for higher fidelity, have made remarkable progress, these two problems have traditionally been treated as separate tasks.

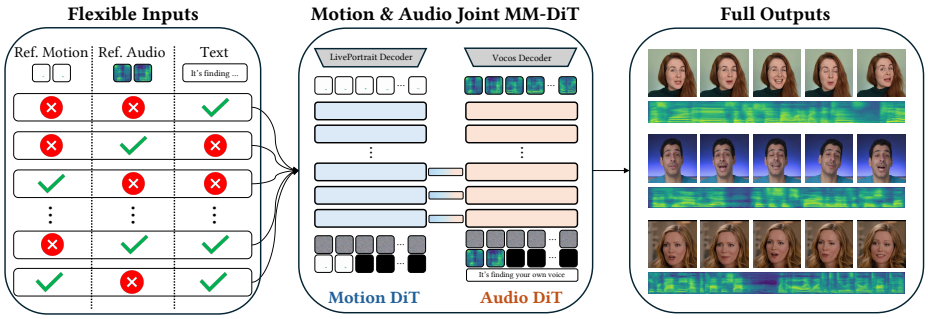

Fig. 1: Overview of our JAM-Flow framework for flexible and joint generation of facial motion and speech. The model accepts diverse input combinations, including text, reference motion, and reference audio. These are processed by our novel Motion & Audio Joint MM-DiT, which enables synchronized synthesis of full audio-visual outputs supporting tasks like talking head generation from text, audio-driven animation, and cross-modal reconstruction (e.g., audio from motion).

Yet in natural human communication, not only is real-time-like latency important, but facial motion and speech are also deeply interwoven. Movements of the mouth, cheeks, and jaw are not merely visual artifacts but integral components of spoken language. Surprisingly, despite this intrinsic connection, few prior works have jointly addressed talking head generation and speech synthesis in a unified model. Existing talking head models typically treat audio as a unidirectional condition, while TTS systems remain blind to facial dynamics.

In this work, we introduce the first training framework that can simultaneously model, generate, and condition on both audio and facial motion modalities within a single flow matching-based framework, while supporting generation at near real-time latency. We integrate two specialized flow matching models: a Motion-DiT for generating implicit facial keypoint sequences, and an Audio-DiT for denoising mel-spectrograms from text and reference speech. These modules are coupled through selective joint attention layers, where only half the layers are fused, allowing effective cross-modal communication while preserving the benefits of modality-specific representations. Furthermore, to inject structural inductive biases into the model, we incorporate rotary positional embeddings (RoPE) [51] with task-specific engineering improvements and attention masking to restrict temporal receptive fields.

One of our key experimental observations is that models should be initialized from well-trained components in their respective modalities before learning cross-

modal relations. In our framework, we initialize the TTS branch with a pretrained F5-TTS [3] model, while the Motion-DiT is first trained separately using the LivePortrait [19] framework to capture facial motion via compact keypoints. After this modality-specific pretraining, the two modules are jointly trained with shared attention layers and an inpainting-style supervision scheme. We propose that this joint inpainting-style supervision is particularly well-suited for optimizing pretrained unimodal models to learn robust cross-modal interactions, enabling flexible generation even under partially missing inputs.

Our joint inpainting-style supervision not only facilitates effective cross-modal learning but also enables versatile inference. By design, each modality can be conditioned on full, partial, or even absent inputs, allowing a single model to flexibly support diverse tasks and generation scenarios. As illustrated in Figure 1, our model supports a wide range of input-output configurations, including talking head generation, TTS, and cross-modal reconstruction, within a single coherent framework. To the best of our knowledge, this is the first attempt to employ joint inpainting-style supervision to achieve such flexible cross-modal synthesis, marking an important step toward practical and scalable multimodal generation.

Our contributions are summarized as follows:

1. We analyze and leverage pretrained implicit representations to enable talking head video generation and TTS synthesis at near real-time latency.
2. We present the first inpainting-style joint training framework for talking head and TTS generation, which integrates both the model architecture and training strategy to enable mutual conditioning across modalities.
3. We demonstrate that our inpainting-style joint supervision effectively learns cross-modal relations, allowing robust and versatile generation across diverse tasks.
4. We further provide a practical architectural design for connecting pretrained unimodal models, including partial joint attention, RoPE integration, and attention masking, tailored for multi-modal flow matching.

2 Related Work

2.1 Flow Matching and Multimodal Diffusion Transformer

Flow matching [12, 33, 34] enhances generative modeling efficiency over score-based diffusion models [23, 50]. It learns a continuous transformation Z_t between distributions via an Ordinary Differential Equation (ODE): $dZ_t/dt = v_\theta(Z_t, t)$, where v_θ is a learnable vector field. The objective is to match v_θ to a target velocity field $u_t(x)$. Conditional Flow Matching (CFM) [33] and Rectified Flow (RF) [34] simplify this: for paired samples $x_0 \sim \pi_0$ and $x_1 \sim \pi_1$, they define an intermediate point via linear interpolation, $x_t = (1-t)x_0 + tx_1$, and set the target velocity $u_t(x_t)$ to be the difference $x_1 - x_0$. This results in the CFM loss:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,x_0,x_1}\left[\|v_\theta(x_t, t) - (x_1 - x_0)\|^2\right]. \tag{1}$$

The *Multimodal Diffusion Transformer* (MM-DiT) [12] builds on flow matching for joint multi-modal generation. It uses separate DiT [39] branches per modality, fused via joint self-attention for cross-modal interaction, enabling scalable and expressive generation. Large-scale MM-DiTs like Stable Diffusion 3 [12], Flux [28], and CogVideoX [72] achieve SOTA in various tasks (e.g., text-to-image, video synthesis), but none have explored inpainting-style joint training to simultaneously synthesize two modalities.

2.2 Talking Head Generation

Talking head generation synthesizes realistic facial animations from conditional signals, mainly through video-driven (facial reenactment) or audio-driven methods. Video-driven approaches [19, 49, 67] animate a source portrait using motion cues from a driving video. Early methods often used facial landmarks [49] or 3D models [55], while recent works like LivePortrait [19] use implicit keypoints and X-portrait [67] employs diffusion control. In contrast, audio-driven methods [41, 57, 69, 80] create lip movements synchronized with input audio. Pioneering work like Wav2Lip [41] focused on lip-sync accuracy with GAN, while later methods (e.g., EMO [57], Omni-Human [32]) often use diffusion models for enhanced expressiveness.

Current methods [1, 2, 8, 13, 14, 17, 25, 31, 35, 46, 56, 58, 62, 65, 71, 76] typically model a uni-directional audio-to-visual flow based on pre-trained large-scale text-to-video diffusion transformers. However, facial motion and speech are mutually influential in real conversations. Our hybrid approach addresses this by combining joint diffusion-based generation of keypoints and audio with efficient pixel decoding via LivePortrait, enabling bidirectional information exchange and flexible generation.

2.3 Neural Text-to-Speech Generation

Neural text-to-speech (TTS) has progressed from attention-based sequence-to-sequence models [64] to more advanced architectures. Early non-autoregressive systems [42, 43] enhanced efficiency via explicit duration modeling. A significant shift towards zero-shot voice cloning was pioneered by neural codec language models [60], which autoregressively generate speech tokens from minimal reference audio, despite some inference challenges.

Diffusion-based methods have since emerged as a powerful paradigm, particularly for zero-shot voice cloning. These include approaches demonstrating high-quality speech via diffusion [27, 29, 47, 54] and flow matching for efficient text-guided speech infilling [3]. Recent efforts focus on refining speech-text alignment and moving from explicit duration to more flexible frameworks [11, 30]. F5-TTS [3] advances this with flow matching and Diffusion Transformers (DiTs) for efficient non-autoregressive generation. While research continues to address alignment robustness, prosody, and efficiency [26], the simultaneous generation of speech and matching lip motion remains largely unaddressed.

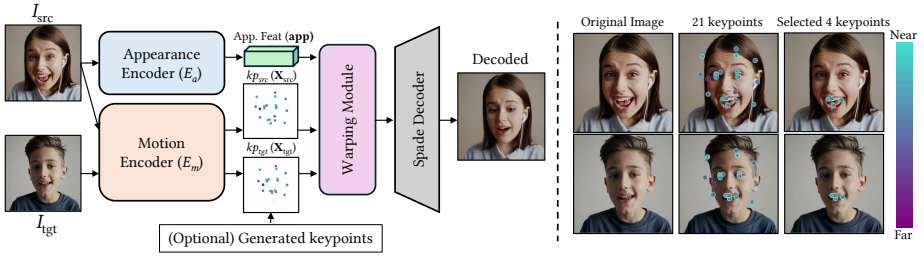

Fig. 2: LivePortrait framework and mouth-related expression keypoint analysis. LivePortrait's motion encoder (E_m) infers parameters including a 3D expression deformation $\mathbf{e} \in \mathbb{R}^{21\times3}$ for 21 canonical keypoints. We find that deforming approximately four specific keypoints (highlighted) primarily dictates mouth articulation. Our Motion-DiT leverages this by generating only the deformation components ($\mathbf{e}^{\text{mouth}} \subset \mathbf{e}$) for these crucial mouth keypoints, enabling efficient lip-sync.

2.4 Automated Video Dubbing

Automated video dubbing synthesizes speech from text and video inputs, focusing on aligning generated speech with existing visual content, unlike video generation from audio/text. It extends text-to-speech (TTS) by conditioning on visual context, especially lip movements and facial expressions. The main challenge is ensuring temporal synchronization and reflecting visual expressiveness in the synthesized speech. Prior work [5, 6, 52] has explored aligning visual cues, multi-scale style learning, and audio-visual fusion. Notably, our model, though not explicitly optimized for this task, shows an emergent capability for generating speech well-aligned with lip movements in given videos.

3 Preliminaries

3.1 LivePortrait Framework

LivePortrait [19] enables image-to-video synthesis by disentangling structural and appearance features from a single input image. It employs an appearance encoder E_a to extract a global appearance feature $\mathbf{app} \in \mathbb{R}^{C\times D'\times H'\times W'}$. In parallel, a motion encoder E_m infers a set (21) of 3D implicit facial keypoints $\mathbf{X} \in \mathbb{R}^{21\times3}$ which is parametrized by canonical keypoints $\mathbf{x}_c \in \mathbb{R}^{21\times3}$, pose matrix $\mathbf{R} \in \mathbb{R}^{3\times3}$, expression deformation $\mathbf{e} \in \mathbb{R}^{21\times3}$, scale $\mathbf{s} \in \mathbb{R}$ and translation $\mathbf{t} \in \mathbb{R}^{21\times3}$. The final keypoints are then computed as: $\mathbf{X} = \mathbf{s} \cdot (\mathbf{x}_c\mathbf{R} + \mathbf{e}) + \mathbf{t}$. The warping module modifies $\mathbf{app}$ with estimated keypoint differences and the decoder projects this warped appearance feature $\mathbf{app}'$ into a target video frame.

We analyze this implicit embedding to identify its functional structure. Visualizing the 21 dimensions of the expression embedding $\mathbf{e}$ (Figure 2) reveals that approximately four specific dimensions consistently control the mouth region. Isolating these mouth-related components $\mathbf{e}^{\text{mouth}} \subset \mathbf{e}$ and freezing others

modifies only the lip shape. This empirical finding suggests lip-sync generation can be simplified by modeling only this small subset of expression dimensions. We leverage this by training our motion generation module to predict only these mouth-related components, which are then combined with fixed identity- and pose-related features for rendering. Importantly, this implicit representation exposes a low-dimensional subspace corresponding to facial expressions and lip shapes within an otherwise high-dimensional facial motion space, which plays a crucial role in enabling near real-time sampling.

3.2 F5-TTS and Conditional Flow Matching

F5-TTS [3] is a Conditional Flow Matching (CFM) based text-to-speech model. Inspired by inpainting-based approaches such as VoiceBox [29], F5-TTS treats speech generation as a mask-and-predict problem, where parts of the mel-spectrogram are randomly masked during training and then reconstructed conditioned on surrounding context and reference signals.

Formally, the model receives a masked audio segment $\tilde{\mathbf{a}}^{\text{masked}}$ and a conditioning vector $\mathbf{c}^{\text{text}}$ consisting of unmasked regions, reference audio features, and text embeddings. These are concatenated and passed through a flow-based network trained using the CFM objective described in Eq. 1.

This inpainting formulation has two key benefits. First, it enables robust training across diverse conditioning scenarios by randomly varying the masked regions. Second, it inherently models the relationship between unmasked context and the regions to be reconstructed, ensuring consistent speech generation. As a result, F5-TTS produces high-quality, voice-consistent speech for arbitrary text inputs and provides a strong foundation for our Audio-DiT module.

4 Method

4.1 Overview

Our goal is to simultaneously generate temporally aligned speech audio and facial motion from multimodal inputs (text, reference audio, or motion). To this end, we propose a dual-stream diffusion architecture composed of Audio-DiT and Motion-DiT, partially fused via joint attention blocks. Figure 3 illustrates the overall architecture with training and inference pipeline.

4.2 Motion-DiT and Audio-DiT Design and Flow Matching

The Motion-DiT generates expression embeddings $\mathbf{e}^{\text{mouth}} \in \mathbb{R}^{T_{\text{frame}} \times 4 \times 3}$ that control lip motion, following the observation from Section 3.1 that 4 of the 21 expression dimensions are sufficient to model mouth dynamics. As shown in Figure 3, we provide two inputs to the motion stream: the target expression excluding mouth-related components, denoted $\mathbf{e}^{\text{rest}} \in \mathbb{R}^{T_{\text{frame}} \times 17 \times 3}$ and the audio-derived conditioning features $\mathbf{f}_{\text{audio}}$ from the Audio-DiT stream (if available). During

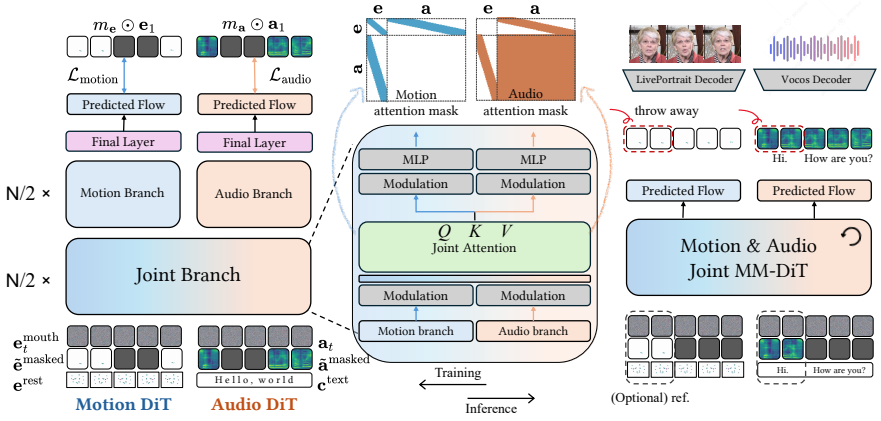

Fig. 3: The training and inference pipeline of the JAM-Flow framework. Our joint MM-DiT comprises a Motion-DiT for facial expression keypoints ($\mathbf{e}^{\text{mouth}}$) and an Audio-DiT for mel-spectrograms ($\mathbf{a}$), coupled via joint attention. The model is trained with an inpainting-style flow matching objective on masked inputs and various conditions (text, reference audio/motion). At inference, it flexibly generates synchronized audio-visual outputs from partial inputs.

training, the Motion-DiT denoises corrupted $\mathbf{e}^{\text{mouth}}$ vectors via a conditional flow-matching process:

$$\mathcal{L}_{\text{motion}} = \mathbb{E}_{\mathbf{e}_0^{\text{mouth}}, \mathbf{e}_1^{\text{mouth}}, t} \left[\| v_\theta(\mathbf{e}_t^{\text{mouth}}, t; \mathbf{f}^{\text{audio}}, \tilde{\mathbf{e}}^{\text{masked}}, \mathbf{e}^{\text{rest}}) - (\mathbf{e}_1^{\text{mouth}} - \mathbf{e}_0^{\text{mouth}}) \|^2 \right] \tag{2}$$

where $\mathbf{e}_t$ denotes the noisy input at step t, and v_θ is the velocity predicting DiT. This design naturally enables joint generation of two tightly related modalities, while representing lip motion in a low-dimensional space, which is key to efficient and near real-time synthesis.

The Audio-DiT generates mel-spectrograms $\mathbf{a}_{\text{mel}} \in \mathbb{R}^{T_{\text{mel}} \times d_{\text{mel}}}$ using the Conditional Flow Matching (CFM) objective. Following F5-TTS, we apply random inpainting masks to audio segments and condition on reference audio and text features. Additionally, we adapt the motion-derived conditioning features $\mathbf{f}^{\text{motion}}$ from the Motion-DiT stream (if available). As defined in Eq. 1, the model learns to reconstruct missing regions from context:

$$\mathcal{L}_{\text{audio}} = \mathbb{E}_{t, \mathbf{a}_0, \mathbf{a}_1} \left[\| v_\theta(\mathbf{a}_t, t; \mathbf{f}^{\text{motion}}, \tilde{\mathbf{a}}^{\text{masked}}, \mathbf{c}^{\text{text}}) - (\mathbf{a}_1 - \mathbf{a}_0) \|^2 \right] \tag{3}$$

This enables the Audio-DiT to flexibly operate under partial or complete conditions and naturally generalize to reference audio-based speech generation.

The final flow-matching loss is defined as

$$\mathcal{L} = \mathcal{L}_{\text{audio}} + \mathcal{L}_{\text{motion}}. \tag{4}$$

During training, audio and motion streams are randomly masked: sometimes only motion is hidden, sometimes only audio, and at other times both are missing.

This stochastic masking compels the model to attend jointly to the available context across modalities, naturally learning dependencies between unmasked and masked regions.

4.3 Joint Attention and Temporal Fusion

To enable cross-modal interactions, we introduce N_{joint} layers of joint attention between the audio and motion streams, as illustrated in Figure 3. Tokens from both streams are fused via joint-attention blocks, while the rest of the transformer layers remain modality-specific.

To ensure temporal compatibility, we apply a simple alignment of rotary positional embeddings (RoPE). For position index $p \in [0, L-1]$ and frequency θ_d, the rotation angle is defined as $\phi(p, d) = \frac{p}{L} L_{\text{ref}} \cdot \theta_d$, where L is the sequence length of the current modality and $L_{\text{ref}} = \max(L_a, L_m)$. This scaling ensures that audio and motion tokens at the same timestep share comparable angular positions, despite having different sequence lengths. While this adjustment is straightforward, we found it to be indispensable: without such alignment, joint attention fails to converge, as queries and keys from different modalities interact at incompatible positional scales.

4.4 Attention Masking Strategy

A key novelty of our framework lies in the attention masking design for joint modeling. Unlike prior multimodal transformers (e.g., MM-DiT) that apply symmetric attention across streams, we introduce modality-specific asymmetric masks tailored to the functional roles of audio and motion.

For motion tokens, we impose a local self-attention window, reflecting the inductive bias that facial motion depends mainly on temporally adjacent frames. Motion-to-audio attention is restricted to the corresponding local window, while audio-to-audio attention is disabled during motion generation. Conversely, for audio tokens, we retain full global self-attention across the utterance, but restrict cross-modal attention to only motion tokens at the same timestamp, with motion-to-motion self-attention entirely masked out. (Figure 3, top)

This design is simple yet crucial: each modality attends in the way most natural to its generative process, while their cross-modal interactions remain tightly aligned. To our knowledge, no prior work has explored such asymmetric, modality-specific masking for joint attention. We found this strategy indispensable for stable training and synchronized outputs.

4.5 Training Procedure

Training is conducted in two stages:

Stage 1: We prepare pretrained models for each modality. For audio, we adopt F5-TTS as the pretrained TTS backbone. For motion, we train a model from scratch following a common practice in talking-head generation, where wav2vec2 features are concatenated to the motion input.

Table 1: Talking-head generation comparison on HDTF [78]. **Inference** is measured on a **single RTX A6000** for generating a **20-second** video, unless otherwise noted. †SadTalker is reported with GFPGAN. ‡Hallo3 is reported on an **H100** GPU. With additional engineering and an H100-class GPU, our method can approach real-time generation (19s for a 20s video at **25 fps**).

Method	FID ↓	FVD ↓	LSE-C ↑	LSE-D ↓	Inference (20s) ↓	#Params
Ground Truth	-	-	8.70	6.597	-	-
SadTalker [77]	22.340	203.860	7.885	7.545	2.5 min†	∼200M
DreamTalk [36]	78.147	890.660	6.376	8.364	-	-
AniPortrait [66]	26.561	234.666	4.015	10.548	7 min	∼1B
Hallo [68]	20.545	173.497	7.750	7.659	23 min	1–3B
Hallo3 [7]	20.359	160.838	7.252	8.106	30 min‡	1–3B
Ours-Stage1	18.372	194.27	7.138	7.947	-	-
Ours (I2V)	17.571	192.30	7.324	7.777	45 sec	∼500M (joint)
Ours (V2V)	**11.633**	**25.07**	**8.086**	**7.181**	45 sec	∼500M (joint)

Stage 2: We then connect the two modalities with joint attention layers, applying RoPE alignment to ensure compatible temporal embeddings. Both branches are jointly trained with gradients flowing across modalities through the shared attention layers, enabling mutual refinement of audio and motion representations. At this stage, wav2vec2 features are no longer used: text conditioning is injected via the Audio-DiT, while the Motion-DiT consumes both $\mathbf{e}_{\text{rest}}$ and intermediate audio features $\mathbf{f}^{\text{audio}}$. Further architectural and training details are provided in the supplementary material.

5 Experiments

5.1 Experimental Setup

We train our model on the CelebV-Dub [52] dataset, filtered from CelebV-HQ [81] and CelebV-Text [75]. Training is conducted in two stages: the Motion-DiT is first trained from scratch using keypoint representations, while the Audio-DiT is initialized from F5-TTS. After convergence, joint training is performed on paired audio-visual data. Evaluation is carried out on CelebV-Dub test splits and HDTF [78], using standard metrics including WER, SIM-o, LSE-C, LSE-D, spkSIM, FID, and FVD. During our experiments, we found that CelebV-Dub is annotated with *Whisper-generated pseudo-transcripts* rather than ground-truth transcripts, introducing nontrivial label noise; as a result, performance can be slightly lower than that of dedicated TTS models trained on clean GT pairs.

5.2 Quantitative Evaluation

Our model is the first to be explicitly trained for joint audio–motion generation, and thus there are no direct baselines for this unified setting. Nevertheless, by

controlling which inputs remain unmasked at inference, the same model can flexibly perform multiple tasks, for example, audio-only, motion-only, or full audiovisual generation. To demonstrate its versatility, we evaluate across three representative tasks by comparing with models that are individually specialized for each, emphasizing that our approach does not rely on task-specific architectures yet still achieves competitive results. Because common objective metrics may not fully reflect perceptual quality and synchronization—especially under codec variations or TTS-generated audio—we additionally report a user study to better capture human preferences in the supplementary material.

Talking head generation. We evaluate talking head performance on HDTF using four key metrics: FID, FVD, LSE-C, and LSE-D. As shown in Table 1, our model performs competitively or better with SOTA methods such as SadTalker, AniPortrait, and Hallo3. Notably, our method is primarily designed with a video-to-video (V2V) setup in mind, utilizing a sequence of 17 non-mouth expression keypoints as a 'clue' (following [79]). This keypoint 'clue' is used in both V2V and image-to-video (I2V) configurations: V2V warps frames sequentially from a source video, whereas I2V consistently warps an initial source image. In addition to quality, our approach is substantially faster: on a single RTX A6000, it generates a 20-second video in 45 seconds, and with additional engineering on an H100-class GPU it can approach real-time throughput (e.g., 19 s for 20 s at 25 fps; see Table 1). Moreover, this quantitative improvement aligns with human perception: in our user study on HDTF, participants ranked our V2V and I2V variants as the top two methods in overall quality (see the user study section in the supplementary material). Notably, we did not adopt any explicit diffusion distillation [24, 48, 73, 74], and we believe there is still room to make it faster.

Table 2: Comparison of text-to-speech performance in LibriSpeech-PC test-clean [3, 38] benchmark. Methods are from prior work [3, 10, 11, 18, 26]

Method	**WER ↓**	**SIM-o ↑**
Cosy Voice	3.59%	0.66
FireRedTTS	2.69%	0.47
E2 TTS	2.95%	0.69
F5-TTS	2.42%	0.66
MegaTTS 3	**2.31%**	**0.70**
Ours	4.91%	0.64

Text-to-speech generation. TTS performance is evaluated using WER and SIM-o on LibriSpeech-PC test-clean [3, 38]. Although our model is trained for joint audio–motion generation, for a fair comparison against dedicated TTS systems we perform TTS evaluation by generating *audio only* at inference time. As shown in Table 2, our WER is slightly higher than that of dedicated TTS systems. We stress, however, that our model is not optimized as a pure TTS system; it is a unified audio–motion generator trained under a substantially noisier supervision regime, where WER is largely *data-bounded* rather than *capacity-bounded*.

Concretely, CelebV-Dub provides no human-verified transcripts, and we therefore rely on Whisper-generated pseudo-GT captions. Transcription errors in Whisper directly propagate to the supervision signal and impose a lower bound

Table 3: Comparison of automated video dubbing performance in CelebV-Dub [52] dataset. Methods are from prior work [5, 6, 40, 52]

Method	LSE-C ↑	LSE-D ↓	WER ↓	spkSIM ↑
Ground Truth	6.73	7.44	4.15%	-
Zero-Shot TTS	2.78	11.68	3.83%	0.316
HPMDubbing	**6.36**	**7.80**	24.06%	0.146
StyleDubber	3.78	10.40	9.48%	0.264
VoiceCraft-Dub	6.05	8.33	7.01%	0.333
Ours	3.43	10.56	**6.39%**	**0.410**

on attainable WER. This is consistent with prior findings on LibriSpeech-PC: the reported WER of Whisper-base is 4.50% [3], which closely matches the ∼4.9% WER we observe, indicating that label noise is the primary bottleneck. Moreover, the audio track in CelebV-Dub is obtained via source separation (Spleeter [22]), which introduces artifacts absent in standard audio-only TTS corpora and further limits achievable fidelity.

Finally, we caution that WER is not a reliable proxy for perceptual quality, especially in our multimodal setting: small transcript-level errors may have limited impact on intelligibility or speaker similarity, and recent work has highlighted the mismatch between WER and human judgments of transcript readability and severity of errors [45]. Taken together, these factors explain the modest WER gap to specialized TTS baselines, while our qualitative results demonstrate that joint training captures cross-modal relationships without a meaningful degradation in perceived audio quality.

Automated video dubbing. Thanks to the inpainting-based training paradigm, our model naturally extends to automated video dubbing, generating speech that is both semantically correct and temporally aligned with the speaker's lip movements. However, as discussed by [37, 52, 70], we found that SyncNet-derived metrics [4] (LSE-C, LSE-D) often fail to operate reliably, showing instability under codec variations and, most critically, with TTS outputs from our model. We therefore caution against over-interpreting these scores and refer readers to the supplementary materials for qualitative examples that better reflect performance. As seen in Table 3, our model achieves strong WER and the highest spkSIM among all methods, demonstrating its effectiveness for this task. To complement these imperfect objective metrics, we also conduct a user study on CelebV-Dub, where our method is preferred in 62.6% of cases over VoiceCraft-Dub (37.4%), while other baselines receive no votes.

5.3 Qualitative Results

We provide a supplementary webpage with videos. On standard settings (talking-head comparison, TTS comparison, and automated dubbing), our method shows

Table 4: Ablation study on the number of joint attention blocks, motion attention masking, and Audio-DiT finetuning.

# Joint Blocks	Attn. Mask	Train Audio-DiT	FID ↓	LSE-C ↑	LSE-D ↓	WER ↓	SIM-o ↑
22 (Full)	✗	✗	5.759	4.81	8.40	6.88%	0.62
11 (Half)	✗	✗	5.735	4.64	9.07	7.25%	0.62
11 (Half)	✗	✓	5.748	6.44	7.99	7.93%	0.61
11 (Half)	✓	✗	5.747	5.76	8.22	**6.76%**	0.63
11 (Half)	✓	✓	**5.662**	**6.45**	**7.73**	7.28%	**0.64**

stronger lip–audio alignment and speaker consistency. Please see the supplementary videos for side-by-side examples and details.

Our Exclusive use cases.

1. Text → Audio + Motion: Without any reference audio, the model *jointly* generates speech with randomly sampled speaker identity and facial motion that remain tightly synchronized. This text-only joint generation is a primary target of our training.
2. Text + Reference Audio → Audio + Motion: Given target text and a short voice reference, the model co-generates speech in the reference voice and the matching lip motion. This is likewise a core *joint* audio–motion scenario we explicitly train for.
3. Reference Motion + Target Text → Audio: With the video fixed, the synthesized audio adapts its opening/closure timing to the observed lip motion, even when perfect articulation is impossible, indicating active attention to motion during speech generation.
4. Reference Motion → Audio (no text): Supplying only motion still yields plausible, time-aligned speech, showing an implicit mapping from visual articulation to acoustics under partial conditioning.

These results highlight the controllability and robustness of our unified model under partially missing inputs. Please check the supplementary webpage for qualitative examples.

5.4 Ablation Studies

To assess the impact of our design choices, we conduct controlled ablation experiments along three axes: (1) the degree of joint attention between audio and motion streams, (2) the presence of temporal attention masks in the Motion-DiT, and (3) the finetuning of the Audio-DiT during stage-2 training. Given the aforementioned instability of LSE-C and LSE-D with generated audio, and to better assess the commonly adopted audio-driven talking head setup, we calculate LSE-C and LSE-D in the ablation study using generated motion paired with ground truth (GT) audio. WER and SIM-o are computed using both generated motion and audio, while LSE-C and LSE-D use GT audio for stability.

Joint Attention Configuration. We compared a *Full Joint* configuration (all DiT layers share attention) with our *Half Joint* approach (only earlier layers fused). Although Table 4 shows numerically stronger scores for Full Joint, it was both unstable and computationally much heavier. In practice, we found that Half Joint provided a more reliable trade-off: comparable qualitative performance with significantly lower training cost. Our choice of Half Joint therefore reflects a balance between performance and efficiency, which proved most practical for large-scale experiments.

Attention Masking (and RoPE). Removing temporal attention masks (*No Masking*) leads to sharp drops in lip-sync quality. While the degradation may not always be dramatic in raw scores, subjectively the model often fails to achieve coherent joint training at all, producing temporally drifting or unsynchronized motion. In other words, masking is not just helpful but almost indispensable for stable learning. Similarly, RoPE alignment (Section 4.3) is absolutely critical: without it, joint training does not converge at all, which is why we do not report a corresponding ablation in the table.

Audio-DiT Finetuning. We also tested freezing the Audio-DiT during stage-2 training to preserve the strong F5-TTS prior. As shown in Table 2 and Table 4, this setting slightly improves WER but results in weaker synchronization. Our experiments revealed why: keeping Audio-DiT fixed prevents the system from learning a proper joint audio–motion distribution, leaving Motion-DiT to adapt alone. Allowing Audio-DiT to finetune, by contrast, enables both modalities to co-adapt through the shared attention layers, yielding more synchronized and coherent outputs.

Taken together, our experiments show that Half Joint attention, modality-specific masking with RoPE alignment, and Audio-DiT finetuning are essential not only for quantitative gains but also for stable and efficient joint training that truly learns a shared multimodal distribution.

6 Discussion and Ethics

Our experiments show that partial joint attention and localized temporal masking are key to stable and coherent multimodal generation. While full joint attention yields slightly better scores, it often results in unstable training. Our hybrid design balances cross-modal fusion and modality-specific representations, contributing to both robustness and performance. We further find that the compact LivePortrait keypoint warping is not only computationally efficient but also preserves identity and fine-grained facial details; in contrast, diffusion-based talking-head methods (e.g., Hallo/Hallo3) often exhibit noticeable temporal flicker in longer sequences despite handling complex scenes well.

We also observed that the generated speech often reflects emotional cues from facial motion. For instance, smiling motions tend to produce brighter, higher-pitched voices, despite the absence of explicit emotion supervision. This suggests

the model implicitly aligns emotion across modalities, opening possibilities for expressive and emotionally-aware generation. More broadly, the generated audio exhibits coherent prosody with facial motion (e.g., natural pauses during speech breaks and consistent tone shifts), suggesting that joint training captures cross-modal alignment beyond simple temporal synchronization.

A major strength of our model is its versatility. It supports diverse input configurations (e.g., audio-only, motion-only, text + portrait) within a single framework, unlike prior models limited to either talking head synthesis or TTS. This flexibility enables applications such as expressive dubbing, silent video revoicing, and adaptive avatar generation. Interestingly, during simultaneous audio and motion generation, we observe an asymmetric adaptation: the model tends to preserve visual consistency with minimal changes to lip motion, while adjusting the audio more substantially to match the given motion and achieve accurate lip-sync. This behavior likely reflects the relative difficulty of modifying realistic visual motion compared to audio synthesis, and provides a practical mechanism for maintaining visual identity while improving synchronization.

Beyond our current scope, we found that generating two modalities jointly is remarkably effective, and we believe this paradigm can naturally extend to other modality pairs (e.g., depth + video, audio + video). While recent concurrent works have started to explore related directions in native multimodal generation, particularly in the video-audio domain [16, 20, 61], our findings further highlight the effectiveness of inpainting-style joint supervision as a simple and scalable mechanism for learning cross-modal relationships, as evidenced in our joint audio-motion setting. More broadly, we believe this paradigm merits deeper exploration as a general framework for multimodal co-generation.

The JAM-Flow model also raises ethical considerations. While it provides benefits for accessibility, avatars, and creative tools, it may also be misused for deepfakes or amplify biases in the data. To address these risks, we plan to explore safeguards such as watermarking and continue reflecting on responsible deployment as the technology evolves.

7 Conclusion

We presented the first joint training framework for speech and facial motion generation, integrating model architecture and training strategy to enable mutual conditioning across modalities. Our unified flow-matching-based design combines modality-specific DiT modules with selectively applied joint attention, RoPE alignment, and attention masking, producing coherent multimodal synthesis without separate pipelines. In addition, by analyzing and leveraging pretrained representation embeddings, we design the system to enable efficient, near real-time sampling.

Through extensive experiments, we demonstrated that inpainting-style joint supervision effectively learns cross-modal relations and supports flexible inference scenarios such as motion-to-audio, audio-to-motion, and full generation from text and portrait image alone, suggesting it as a powerful paradigm for multimodal

co-generation. Finally, we provided a practical architectural pathway for connecting pretrained unimodal models, showing that partial joint attention, RoPE alignment, and localized masking together balance stability, performance, and efficiency, yielding competitive results with specialized state-of-the-art models.

While these findings are promising, current limitations stem largely from data and compute. CelebV-Dub [52], for instance, contains Whisper-generated captions with transcription errors and demuxed audio with artifacts, while LivePortrait constrains motion modeling largely to facial regions. Nevertheless, we believe our results make a compelling case for unified multimodal training. With more curated datasets and stronger video diffusion backbones [21, 59, 72], future work can further extend this framework, unlocking expressive dubbing, controllable avatars, and more natural human–computer interaction.

Acknowledgments

This work was supported by the National Research Foundation of Korea (RS-2026-25497603: Fundamentals of Online 4D Reconstruction with Semantic Viewpoint Control and Generative Priors for Proactive Content Consumption (70%)), IITP grant (RS-2021-II211343: AI Grad. School Program (SNU) (5%), RQT-25-120117: AI Computing Infrastructure Enhancement User Support Program (5%), RS-2026-25546560: Full-Cycle GenAI Talent Development: Safe, Accessible and Content-creating AI with an Interactive Education Platform (20%)).

References

1. Chen, L., Ma, T., Liu, J., Li, B., Chen, Z., Liu, L., He, X., Li, G., He, Q., Wu, Z.: Humo: Human-centric video generation via collaborative multi-modal conditioning. arXiv preprint arXiv:2509.08519 (2025)
2. Chen, Y., Liang, S., Zhou, Z., Huang, Z., Ma, Y., Tang, J., Lin, Q., Zhou, Y., Lu, Q.: Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. arXiv preprint arXiv:2505.20156 (2025)
3. Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K., Chen, X.: F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. arXiv preprint arXiv:2410.06885 (2024)
4. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13. pp. 251–263. Springer (2017)
5. Cong, G., Li, L., Qi, Y., Zha, Z.J., Wu, Q., Wang, W., Jiang, B., Yang, M.H., Huang, Q.: Learning to dub movies via hierarchical prosody models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
6. Cong, G., Qi, Y., Li, L., Beheshti, A., Zhang, Z., Hengel, A.v.d., Yang, M.H., Yan, C., Huang, Q.: Styledubber: towards multi-scale style learning for movie dubbing. arXiv preprint arXiv:2402.12636 (2024)
7. Cui, J., Li, H., Zhan, Y., Shang, H., Cheng, K., Ma, Y., Mu, S., Zhou, H., Wang, J., Zhu, S.: Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21086–21095 (2025)
8. Ding, Y., Liu, J., Zhang, W., Wang, Z., Hu, W., Cui, L., Lao, M., Shao, Y., Liu, H., Li, X., et al.: Kling-avatar: Grounding multimodal instructions for cascaded long-duration avatar animation synthesis. arXiv preprint arXiv:2509.09595 (2025)
9. Drobyshev, N., Casademunt, A.B., Vougioukas, K., Landgraf, Z., Petridis, S., Pantic, M.: Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
10. Du, Z., Chen, Q., Zhang, S., Hu, K., Lu, H., Yang, Y., Hu, H., Zheng, S., Gu, Y., Ma, Z., et al.: Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. arXiv preprint arXiv:2407.05407 (2024)
11. Eskimez, S.E., Wang, X., Thakker, M., Li, C., Tsai, C.H., Xiao, Z., Yang, H., Zhu, Z., Tang, M., Tan, X., et al.: E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In: 2024 IEEE Spoken Language Technology Workshop (SLT). pp. 682–689. IEEE (2024)
12. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: International Conference on Machine Learning (ICML) (2024)
13. Gan, Q., Yang, R., Zhu, J., Xue, S., Hoi, S.: Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation. arXiv preprint arXiv:2506.18866 (2025)
14. Gao, X., Hu, L., Hu, S., Huang, M., Ji, C., Meng, D., Qi, J., Qiao, P., Shen, Z., Song, Y., et al.: Wan-s2v: Audio-driven cinematic video generation. arXiv preprint arXiv:2508.18621 (2025)
15. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)

16. Google DeepMind: Veo 3. https://deepmind.google/models/veo/ (2026), accessed: 2026-03-01
17. Guan, J., Wang, K., Xu, Z., Yang, Q., Sun, Y., He, S., Liang, B., Cao, Y., Li, Y., Feng, H., et al.: Audcast: Audio-driven human video generation by cascaded diffusion transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10678–10689 (2025)
18. Guo, H.H., Hu, Y., Liu, K., Shen, F.Y., Tang, X., Wu, Y.C., Xie, F.L., Xie, K., Xu, K.T.: Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications. arXiv preprint arXiv:2409.03283 (2024)
19. Guo, J., Zhang, D., Liu, X., Zhong, Z., Zhang, Y., Wan, P., Zhang, D.: Liveportrait: Efficient portrait animation with stitching and retargeting control. arXiv preprint arXiv:2407.03168 (2024)
20. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., et al.: Ltx-2: Efficient joint audio-visual foundation model. arXiv preprint arXiv:2601.03233 (2026)
21. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
22. Hennequin, R., Khlif, A., Voituret, F., Moussallam, M.: Spleeter: a fast and efficient music source separation tool with pre-trained models. Journal of Open Source Software **5**(50), 2154 (2020)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Conference on Neural Information Processing Systems (NeurIPS) (2020)
24. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
25. Jiang, J., Zeng, W., Zheng, Z., Yang, J., Liang, C., Liao, W., Liang, H., Zhang, Y., Gao, M.: Omnihuman-1.5: Instilling an active mind in avatars via cognitive simulation. arXiv preprint arXiv:2508.19209 (2025)
26. Jiang, Z., Ren, Y., Li, R., Ji, S., Zhang, B., Ye, Z., Zhang, C., Jionghao, B., Yang, X., Zuo, J., et al.: Megatts 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis. arXiv preprint arXiv:2502.18924 (2025)
27. Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., Liu, Y., Leng, Y., Song, K., Tang, S., et al.: Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. arXiv preprint arXiv:2403.03100 (2024)
28. Labs, B.F.: Flux.1. https://blackforestlabs.ai/announcing-black-forest-labs/ (2024), accessed: November 2024
29. Le, M., Vyas, A., Shi, B., Karrer, B., Sarı, L., Moritz, R., Williamson, M., Manohar, V., Adi, Y., Mahadeokar, J., Hsu, W.N.: Voicebox: Text-guided multilingual universal speech generation at scale. In: Conference on Neural Information Processing Systems (NeurIPS) (2023)
30. Lee, K., Kim, D.W., Kim, J., Cho, J.: Ditto-tts: Efficient and scalable zero-shot text-to-speech with diffusion transformer. arXiv preprint arXiv:2406.11427 **1** (2024)
31. Li, X., Xie, P., Ren, Y., Gan, Q., Zhang, C., Kong, F., Yin, X., Peng, B., Yuan, Z.: Infinityhuman: Towards long-term audio-driven human. arXiv preprint arXiv:2508.20210 (2025)
32. Lin, G., Jiang, J., Yang, J., Zheng, Z., Liang, C.: OmniHuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. arXiv preprint arXiv:2502.01061 (2025)

33. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow Matching for Generative Modeling. In: International Conference on Learning Representations (ICLR) (2023)
34. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
35. Ma, X., Huang, S., Cai, J., Guan, Y., Zheng, S., Zhao, H., Zhang, Q., Zhang, S.: Playmate2: Training-free multi-character audio-driven animation via diffusion transformer with reward feedback. arXiv preprint arXiv:2510.12089 (2025)
36. Ma, Y., Zhang, S., Wang, J., Wang, X., Zhang, Y., Deng, Z.: Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. arXiv preprint arXiv:2312.09767 **2**(3) (2023)
37. Muaz, U., Jang, W., Tripathi, R., Mani, S., Ouyang, W., Gadde, R.T., Gecer, B., Elizondo, S., Madad, R., Nair, N.: Sidgan: High-resolution dubbed video generation via shift-invariant learning. In: IEEE International Conference on Computer Vision (ICCV) (2023)
38. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an asr corpus based on public domain audio books. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 5206–5210. IEEE (2015)
39. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: IEEE International Conference on Computer Vision (ICCV) (2023)
40. Peng, P., Huang, P.Y., Li, S.W., Mohamed, A., Harwath, D.: Voicecraft: Zero-shot speech editing and text-to-speech in the wild. arXiv preprint arXiv:2403.16973 (2024)
41. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: 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. pp. 484–492 (2020)
42. Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.Y.: Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558 (2020)
43. Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.Y.: Fastspeech: Fast, robust and controllable text to speech. Conference on Neural Information Processing Systems (NeurIPS) (2019)
44. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
45. Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.L., et al.: English conversational telephone speech recognition by humans and machines. arXiv preprint arXiv:1703.02136 (2017)
46. Seo, J., Mira, R., Haliassos, A., Bounareli, S., Chen, H., Tran, L., Kim, S., Landgraf, Z., Shen, J.: Lookahead anchoring: Preserving character identity in audio-driven human animation. arXiv preprint arXiv:2510.23581 (2025)
47. Shen, K., Ju, Z., Tan, X., Liu, Y., Leng, Y., He, L., Qin, T., Zhao, S., Bian, J.: Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. arXiv preprint arXiv:2304.09116 (2023)
48. Shin, J., Li, Z., Zhang, R., Zhu, J.Y., Park, J., Shechtman, E., Huang, X.: Motionstream: Real-time video generation with interactive motion controls. arXiv preprint arXiv:2511.01266 (2025)
49. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Conference on Neural Information Processing Systems (NeurIPS) (2019)

50. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-Based Generative Modeling through Stochastic Differential Equations. In: International Conference on Learning Representations (ICLR) (2021)
51. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
52. Sung-Bin, K., Choi, J., Peng, P., Chung, J.S., Oh, T.H., Harwath, D.: VoiceCraft-Dub: Automated Video Dubbing with Neural Codec Language Models. arXiv preprint arXiv:2504.02386 (2025)
53. Tan, S., Ji, B., Bi, M., Pan, Y.: Edtalk: Efficient disentanglement for emotional talking head synthesis. In: European Conference on Computer Vision. pp. 398–416. Springer (2024)
54. Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., et al.: Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(6), 4234–4245 (2024)
55. Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H.P., Pérez, P., Zollhofer, M., Theobalt, C.: Stylerig: Rigging stylegan for 3d control over portrait images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6142–6151 (2020)
56. Tian, L., Hu, S., Wang, Q., Zhang, B., Bo, L.: Emo2: End-effector guided audio-driven avatar video generation. arXiv preprint arXiv:2501.10687 (2025)
57. Tian, L., Wang, Q., Zhang, B., Bo, L.: Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In: European Conference on Computer Vision (ECCV). Springer (2024)
58. Tu, S., Pan, Y., Huang, Y., Han, X., Xing, Z., Dai, Q., Luo, C., Wu, Z., Jiang, Y.G.: Stableavatar: Infinite-length audio-driven avatar video generation. arXiv preprint arXiv:2508.08248 (2025)
59. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
60. Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., et al.: Neural codec language models are zero-shot text to speech synthesizers. arXiv preprint arXiv:2301.02111 (2023)
61. Wang, J., Qiang, C., Guo, Y., Wang, Y., Zeng, X., Zhang, C., Wan, P.: Klear: Unified multi-task audio-video joint generation. arXiv preprint arXiv:2601.04151 (2026)
62. Wang, R., Feng, J., Tian, L., Luo, H., Li, C., Zhou, L., Zhang, H., Wu, Y., He, X.: Joyavatar: Unlocking highly expressive avatars via harmonized text-audio conditioning. arXiv preprint arXiv:2602.00702 (2026)
63. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
64. Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al.: Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135 (2017)
65. Wang, Z., Wang, J., Ma, K., Lin, D., Zhou, B.: Talkverse: Democratizing minute-long audio-driven video generation. arXiv preprint arXiv:2512.14938 (2025)
66. Wei, H., Yang, Z., Wang, Z.: Aniportrait: Audio-driven synthesis of photorealistic portrait animation. arXiv preprint arXiv:2403.17694 (2024)
67. Xie, Y., Xu, H., Song, G., Wang, C., Shi, Y., Luo, L.: X-portrait: Expressive portrait animation with hierarchical motion attention. In: ACM SIGGRAPH (2024)

68. Xu, M., Li, H., Su, Q., Shang, H., Zhang, L., Liu, C., Wang, J., Yao, Y., Zhu, S.: Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. arXiv preprint arXiv:2406.08801 (2024)
69. Xu, S., Chen, G., Guo, Y.X., Yang, J., Li, C., Zang, Z., Zhang, Y., Tong, X., Guo, B.: Vasa-1: Lifelike audio-driven talking faces generated in real time. Conference on Neural Information Processing Systems (NeurIPS) (2024)
70. Yaman, D., Eyiokur, F.I., Bärmann, L., Akti, S., Ekenel, H.K., Waibel, A.: Audio-visual speech representation expert for enhanced talking face video generation and evaluation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop (2024)
71. Yang, S., Kong, Z., Gao, F., Cheng, M., Liu, X., Zhang, Y., Kang, Z., Luo, W., Cai, X., He, R., et al.: Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. arXiv preprint arXiv:2508.14033 (2025)
72. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
73. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T.: Improved Distribution Matching Distillation for Fast Image Synthesis. arXiv preprint arXiv:2405.14867 (2024)
74. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
75. Yu, J., Zhu, H., Jiang, L., Loy, C.C., Cai, W., Wu, W.: Celebv-text: A large-scale facial text-video dataset. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
76. Zhang, J., Zhu, J., Gan, Z., Luo, D., Lin, C., Xu, F., Peng, X., Hu, J., Liu, Y., Hong, Y., et al.: Soul: Breathe life into digital human for high-fidelity long-term multimodal animation. arXiv preprint arXiv:2512.13495 (2025)
77. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: SadTalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
78. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
79. Zhong, W., Fang, C., Cai, Y., Wei, P., Zhao, G., Lin, L., Li, G.: Identity-preserving talking face generation with landmark and appearance priors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
80. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: MakeItTalk: Speaker-aware talking-head animation. In: ACM SIGGRAPH Asia (2020)
81. Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., Loy, C.C.: Celebv-hq: A large-scale video facial attributes dataset. In: European Conference on Computer Vision (ECCV) (2022)