Textile & Leather Review

ISSN 2623-6281 | www.tlr-journal.com | 10.31881/TLR

An Implicit Motion Guided Latent Completion Framework for Audio-Driven Talking Head Synthesis

Xiang Li, Xiaoli Huang, Xiaonan Luo

How to cite: Li X, Huang X, Luo X. An Implicit Motion Guided Latent Completion Framework for Audio-Driven Talking Head Synthesis. Textile & Leather Review. 2026; 9:2607-2628. https://doi.org/10.31881/TLR.2026.2607

How to link: https://doi.org/10.31881/TLR.2026.2607

Published: 25 April 2026

An Implicit Motion Guided Latent Completion Framework for Audio-Driven Talking Head Synthesis

Xiang Li[1], Xiaoli Huang[2]*, Xiaonan Luo[1]*

[1]Guilin University of Electronic Technology, Guillin 541004, Guangxi, China

[2]Nanfang College·Guangzhou, Guangzhou 510970, Guangdong, China

*huangxl@nfu.edu.cn, luoxn@guet.edu.cn

Article

https://doi.org/10.31881/TLR.2026.2607

Published 25 April 2026

ABSTRACT

Audio-driven talking head synthesis requires simultaneous achievement of photorealistic visual quality, accurate lip synchronization, robust identity preservation, and computational efficiency. With the rapid proliferation of digital humans, these capabilities are becoming increasingly critical for widespread applications in human-computer interaction. Methods relying on raw audio conditioning suffer from identity-averaged dynamics, while those using explicit geometric priors (2D landmarks or 3DMM coefficients) face inherent information bottlenecks that cause over-smoothing and loss of fine-grained lip details. We present IMGTalk, an Implicit Motion Guided latent completion framework that addresses these limitations through a principled two-stage decoupled design. In Stage 1, a Transformer-based Audio-to-Implicit Motion Encoder learns a cross-identity mapping without per-subject fine-tuning from HuBERT audio features to implicit expression deformation , conditioned on each subject 's neutral silent state to enable cross-identity adaptation without per-subject fine-tuning. In Stage 2, an Implicit Motion-Conditioned UNet (IMC-UNet) completes the masked mouth region entirely within the VAE latent space, driven by the predicted expression deformation via positional-encoded cross-attention. We further introduce Reference Spatial Cross-Attention, a reference injection mechanism that replaces naive channel-wise concatenation by projecting the reference latent at every U-Net layer, enabling spatially selective retrieval of lip and teeth appearance information. Extensive experiments on the HDTF and TFHP benchmarks demonstrate that IMGTalk achieves state-of-the-art performance across all evaluation metrics.

KEYWORDS

audio-driven talking head, lip synchronization, digital human, human-computer interaction, implicit motion representation

INTRODUCTION

The rapid advancement of generative modeling has catalyzed unprecedented progress in digital human synthesis, with audio-driven talking head generation emerging as one of the most demanding and practically consequential frontiers. Digital humans are poised to revolutionize various sectors, finding extensive applications in human-computer interaction (HCI) By animating a target speaker's face in synchrony with an input speech signal, such systems underpin a wide range of real-world scenarios, including virtual presenters, conversational agents, and immersive telepresence. In these interactive contexts, the ability to generate photorealistic expressions with precise lip synchronization is paramount for establishing user trust and ensuring a natural, engaging communication experience. Despite years of intensive research, simultaneously achieving photorealistic visual quality, accurate lip synchronization, and robust identity preservation while maintaining computational efficiency remains an open challenge.

Early methods in this domain approached the problem through signal processing and rule-based viseme mapping, relying on handcrafted correspondences between phonemes and mouth shapes [1,2]. The emergence of deep generative models, particularly generative adversarial networks (GANs) [3,4] and, more recently, diffusion models [5,6], fundamentally transformed the landscape. End-to-end frameworks such as Wav2Lip [3] demonstrated that training with a dedicated lip-sync discriminator could yield strong audio-visual alignment, while approaches like EMO [5] and Hallo3 [7] leveraged video diffusion transformers to produce highly expressive and temporally coherent animations. Nevertheless, these methods tend to require massive multi-speaker corpora during training, often yielding averaged mouth dynamics that sacrifice speaker-specific articulation patterns. Moreover, the absence of explicit geometric control makes it difficult to faithfully reproduce subtle deformations such as lip-corner tension or jaw dynamics.

To inject greater controllability, a parallel line of research decouples the generation pipeline into two sequential stages: audio-to-motion estimation followed by motion-conditioned rendering. Intermediate representations — including 2D facial landmarks [2], 3D Morphable Model (3DMM) expression coefficients [8,9], and FLAME-based parametric meshes [10] — serve as a bridge between acoustic input and visual output. Methods such as IPLAP [11] exploit the temporal stability of keypoint trajectories to drive lip-synchronized rendering, while frameworks like Facial [12] and EDTalk [13] regress 3DMM parameters from audio features through recurrent networks, enabling person-specific adaptation via fine-tuning on brief target videos. However, the low dimensionality inherent to these intermediate representations introduces an information

bottleneck: high-frequency lip details, micro-expressions, and subtle muscle activations are inevitably lost during compression, resulting in over-smoothed and temporally jittery animations. More recently, latent diffusion-based methods such as MuseTalk [4], liveportrait [14] have demonstrated that operating within the VAE latent space offers advantages for head synthesis: the compressed spatial resolution reduces computational cost, while the pre-trained VAE decoder provides a strong generative prior for reconstructing fine-grained lip and teeth textures. However, MuseTalk relies on audio features as the sole driving condition, without any geometric prior on mouth shape, which causes the network to regress toward identity-averaged dynamics.

In this paper, we present IMGTalk, an implicit motion guided latent completion framework for audio-driven talking head synthesis that addresses these limitations through a two-stage decoupled design. In Stage 1, a Transformer-based Audio-to-Implicit Motion Encoder learns a learns a generalizable, identity-independent mapping from HuBERT audio features to implicit expression deformation, enabling inference on unseen identities. In Stage 2, an Implicit Motion-Conditioned UNet completes the masked mouth region in the VAE latent space, driven by the implicit motion injected via positional-encoded cross-attention. Extensive experiments show that IMGTalk achieves the best performance in visual quality, lip synchronization accuracy, and computational efficiency, validated by both quantitative metrics and a controlled user study.The main contributions of this work are as follows:

A Novel Two-Stage Implicit Motion-Guided Framework: We propose IMGTalk, an audio-driven talking head synthesis framework that decouples the task into a generalizable, training-free cross-identity acoustic-to-implicit-motion mapping and high-fidelity latent space completion. By operating entirely within the VAE latent space and driving generation with disentangled implicit expression offsets rather than raw audio or explicit landmarks, our method achieves high-fidelity texture preservation (e.g., teeth and lip details) while enabling robust zero-shot generalization to unseen identities.

Cross-identity Expression Mapping via Implicit Priors: We introduce a Transformer-based Audio-to-Implicit Motion Encoder that translates HuBERT audio features into implicit expression deformation. By initializing from each subject ‘s neutral silent expression and identity, predicting relative expression offsets with temporal velocity constraints, the encoder enables cross-identity adaptation without any subject-specific fine-tuning at inference time.

Reference Spatial Cross-Attention for Appearance Injection: We propose RSCA, a query-based reference injection mechanism that replaces the naive channel-wise concatenation of reference and source latents. By

projecting the reference VAE latent as Keys and Values and injecting them into every layer of the IMC-UNet via cross-attention, RSCA allows each spatial position at every scale to selectively retrieve appearance information from the reference on demand, leading to more faithful lip texture and teeth structure reconstruction compared to passive concatenation-based baselines.

RELATED WORK

Audio-Driven Talking Head Synthesis

Audio-driven talking head generation has evolved from early rule-based viseme mapping methods into a rich family of deep learning approaches. Broadly, existing methods can be categorized according to whether they target a specific individual (person-specific) or operate across arbitrary identities (generalized).

Person-specific Talking Head

Person-specific methods train or fine-tune models on short video clips of a target individual, allowing them to capture idiosyncratic speech gestures, lip shapes, and facial dynamics unique to that person. Early work required hours of footage to learn reliable audio-lip mappings [15], while subsequent methods progressively reduced this requirement to minutes by leveraging neural rendering paradigms. AD-NeRF [16] pioneered the use of Neural Radiance Fields (NeRF) [17] to model the target speaker as a continuous volumetric scene conditioned on audio features, achieving high-fidelity personalization but at prohibitive inference costs of several seconds per frame. RAD-NeRF [18] improved efficiency through region-aware decomposition and audio-spatial decomposition respectively, yet both remain computationally demanding for real-time use. More recently, 3D Gaussian Splatting (3D-GS) [19] has been introduced into this setting, GaPTalk [9] demonstrated that conditioning a Gaussian deformation field on 3DMM expression and identity parameters enables fine-grained control over lip movements, reducing training time to under one hour while maintaining real-time inference. Despite their personalization advantages, all person-specific methods require per-identity retraining and are not directly applicable to unseen speakers.

Generalized Talking Head

Generalized methods aim to synthesize lip-synchronized talking head videos for arbitrary identities without identity-specific fine-tuning, typically relying on few-shot reference frames at inference time. Wav2lip [3] and IPLAP [11] introduced a robust lip-sync discriminator based on SyncNet to supervise audio-visual alignment, achieving strong synchronization performance across diverse identities. However, its pixel-space generation

tends to produce blurry mouth regions at high resolution, as the network struggles to recover high-frequency textural details(e.g. teeth structure and lip texture) solely from audio-visual correlation loss. DINet [20] addressed this limitation by replacing direct pixel generation with spatial feature deformation, warping reference feature maps to reconstruct the mouth region with improved texture fidelity. VideoReTalking [21] proposes a three-stage pipeline that sequentially normalizes facial expression to a canonical state, performs lip synchronization, and applies face enhancement, achieving improved temporal coherence. MuseTalk [4] moved generation into the VAE latent space, concatenating the masked source and reference latents as UNet input and conditioning on audio features for real-time lip synchronization. Diffusion-based methods such as EMO [5] leverage iterative denoising to generate highly expressive and temporally coherent animations from a single reference image, but their inference processes impose substantial computational costs that preclude real-time deployment.

Intermediate Representation Methods

To bridge the modality gap between audio and facial imagery, a mainstream paradigm decouples talking head generation into two stages: audio-to-motion prediction and motion-conditioned rendering, using intermediate representations as a controllable interface between them.

Landmark-based methods such as IPLAP [11] predict 2D/3D facial keypoints from audio to guide rendering. These methods benefit from the temporal smoothness and geometric interpretability of landmarks, making them relatively easy to train and control. However, landmarks are inherently sparse representations and provide only coarse geometric constraints on lip shape. The discrete keypoint structure cannot encode continuous deformation fields, leaving fine-grained dynamics such as inter-dental visibility, lip-corner puckering, and subtle muscle activations unrepresented.

Parametric model-based methods adopt 3DMM [22] or FLAME [23] for a more structured representation, with frameworks like PIR [8], Facial [12], and EDTalk [13] regressing expression coefficients from audio via recurrent networks. However, the low dimensionality of these parameters creates an information bottleneck: the shared expression space must encode head pose, eye blinks, and all other facial dynamics alongside lip motion, leaving insufficient capacity for high-frequency lip details like lip-corner puckering.

An alternative line of work learns compact implicit keypoints in an unsupervised manner, bypassing the need for predefined geometry. Face vid2vid [24] introduced 3D implicit keypoints that jointly encode head pose and expression deformation. LivePortrait[14] further demonstrated that compact implicit keypoints can function

as a form of implicit blendshapes: by decomposing motion into canonical keypoints, head pose rotation, expression deformation offset, and global scale, the framework achieves disentangled and fine-grained facial control with negligible overhead. In this work, we adopt LivePortrait 's implicit expression feature space as the intermediate representation for audio-driven generation, and operate within the VAE latent space for high-fidelity mouth completion.

METHOD

To address the critical challenges in existing photorealistic talking head synthesis — including limited zero-shot generalization, degraded visual quality in the mouth region, and imprecise audio-lip synchronization — we propose IMGTalk, a novel two-stage implicit motion-guided latent completion framework. As illustrated in Fig.1, IMGTalk decouples the task into two sequential stages: (1) an Audio-to-Implicit Motion Encoder that maps phonetic audio features to implicit expression deformation in a generalizable manner across unseen identities, and (2) an Implicit Motion-Conditioned UNet (IMC-UNet) that completes the masked mouth region in the VAE latent space, guided by the predicted expression and enriched by reference appearance via a Reference Spatial Cross-Attention (RSCA) module.

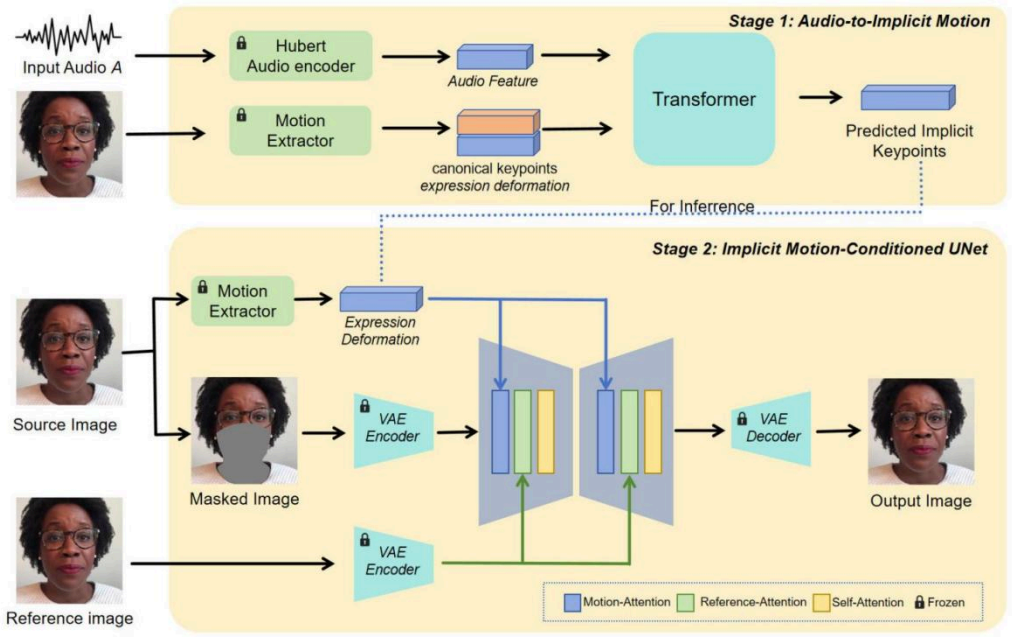

Figure 1. Overview of the IMGTalk framework

Stage 1: A Transformer-based encoder maps HuBERT audio features and a neutral silent-state embedding to implicit expression deformation. Stage 2: The masked source latent $\tilde{z}_s$ is completed by the IMC-UNet, condi-

tioned on $\delta_k^{(t)}$ via cross-attention. A Reference Spatial Cross-Attention (RSCA) module injects the reference latent z_r as Keys and Values at every U-Net layer to guide appearance reconstruction. The output latent $\hat{z}$ is decoded to produce the final frame $\hat{I}$.

Framework Overview

Given a source image I_s of the target speaker and a driving audio sequence A, IMGTalk synthesizes a lip-synchronized output frame $\hat{I}$ while preserving the identity and head pose of the source. A reference image I_r , selected from the same identity with a similar head pose, provides appearance cues for the mouth region completion.

In Stage 1, the Audio-to-Implicit Motion Encoder takes the HuBERT audio features $F_{audio} \in \mathbb{R}^{T\times D}$ and the silent-state embedding e_0 as input, and autoregressively predicts a sequence of expression deformation $\left\{\delta_k^{(t)}\right\}_{t=1}^{T}$ that correspond to the lip dynamics of the driving audio.

In Stage 2, the source image I_s is masked in the mouth region to obtain $\tilde{I}_s$, and both $\tilde{I}_s$ and the reference image I_r are encoded by the frozen VAE encoder $\mathcal{E}$ [25] into latent representations $\tilde{z}_s$ and z_r respectively. The IMC-UNet takes $\tilde{z}_s$ as its primary input and completes the masked mouth latent, conditioned on the predicted expression deformation offsets $\Delta\delta_k^{(t)}$ injected via positional-encoded cross-attention. The reference latent z_r is injected at every layer through the RSCA module to provide appearance guidance. The completed latent $\hat{z}$ is decoded by the frozen VAE decoder $\mathcal{D}$ to produce the output frame $\hat{I} = \mathcal{D}(\hat{z})$.

The two-stage decoupling is motivated by the observation that audio-to-motion mapping and motion-to-appearance rendering impose fundamentally different learning objectives. Stage 1 operates in the structured implicit motion space, where expression deformation is disentangled from identity and head pose, enabling cross-identity generalization without subject-specific fine-tuning. Stage 2 operates in the VAE latent space, where the pre-trained decoder provides a strong generative prior for high-fidelity reconstruction of lip and teeth texture. By passing only the disentangled expression deformation — rather than raw audio features — from Stage1 to Stage 2, the IMC-UNet receives a geometrically meaningful driving signal that directly constrains mouth shape, as opposed to the purely correlation-based audio conditioning used in prior latent inpainting methods such as MuseTalk[4].

Audio to Implicit Motion Encoding

Achieving generalizable audio-driven facial animation across diverse identities remains a fundamental challenge in talking head synthesis. The core difficulty lies in learning a unified audio-to-motion mapping that produces identity-consistent mouth movements for arbitrary unseen subjects, without relying on subject-specific fine-tuning or explicit identity supervision. To address this, we propose the Audio to Implicit Motion Encoding module, which learns a generalizable mapping without per-subject fine-tuning from acoustic input to implicit motion within a unified Transformer-based framework, enabling inference across diverse unseen identities.

Phonetic-Aware Audio Representation.

To achieve high-precision lip-sync and robust generalization across diverse identities, we adopt HuBERT [26] as our primary acoustic backbone. Unlike semantic-heavy models such as Whisper, which are optimized for high-level transcription, HuBERT utilizes self-supervised discrete unit discovery to capture fine-grained phonetic structures. This characteristic is fundamental for modeling the subtle nuances of human articulation. To better align with the implicit motion space, we synchronize these features to a 25 Hz sampling rate, ensuring a one-to-one temporal mapping with the video frames. The resulting audio feature sequence, $F_{audio} \in \mathbb{R}^{T \times D}$ where $D = 1024$, provides an identity-agnostic representation that decouples phonetic content from speaker-specific vocal characteristics, thereby facilitating seamless motion transfer for unseen subjects.

Implicit Motion Representation

To accurately capture the nuanced dynamics of human facial expressions while maintaining identity integrity, we adopt the audio-driven facial motion using the implicit keypoint decomposition introduced in LivePortrait [14], which extends the face-vid2vid [24] framework by introducing an explicit scale factor to resolve the scale-expression entanglement present in the original formulation. Specifically, the motion of a talking head is factorized into four disentangled components: canonical keypoints $x_{c,k} \in \mathbb{R}^{K \times 3}$, a rotation matrix $R \in \mathbb{R}^{3 \times 3}$, a translation vector $t \in \mathbb{R}^3$, a scale factor $s \in \mathbb{R}$, and a per-keypoint expression deformation $\delta_k \in \mathbb{R}^{K \times 3}$. The final image-specific keypoints are computed as:

$$\mathbf{x}_k = s \cdot (\mathbf{x}_{c,k} \cdot R + \delta_k) + \mathbf{t} \tag{1}$$

Unlike traditional explicit methods that rely on pre-defined anatomical landmarks or rigid 3D Morphable Models (3DMMs), this implicit representation facilitates a more flexible and expressive manifold for modeling complex, non-linear facial articulations. Specifically, while 2D facial landmarks offer explicit semantic correspondence, they lack depth information and suffer from geometric ambiguity under non-frontal poses. Although 3DMM expression coefficients provide a normalized cross-subject parameter space conducive to generalization, they fail to capture articulatory details, such as extreme mouth apertures, visible teeth configurations, and speaker-specific lip dynamics.

Identity-Conditioned Motion Encoding

We propose an Identity-Conditioned Motion Encoding mechanism that predicts the implicit keypoints corresponding to the input audio via cross-modal multi-head attention. For a given subject, we first extract its motion parameters from a closed-mouth reference image, obtaining the neutral canonical keypoints x_c and expression deformation $\delta^{neutral}$ as the initial embedding. The audio features extracted from HuBERT are transformed into keys K^A and values V^A , while the motion embedding $\tilde{f}_t$ is projected as queries $Q^{\tilde{F}}$. The attended output is computed as:

$$Att(\mathbf{Q}^{\tilde{F}}, \mathbf{K}^A, \mathbf{V}^A, \mathbf{B}^A) = softmax(\frac{\mathbf{Q}^{\tilde{F}}(\mathbf{K}^A)^T}{\sqrt{d_k}} + \mathbf{B}^A)\mathbf{V}^A \quad (2)$$

where B^A is the alignment bias that restricts each motion frame to attend only to its temporally corresponding audio window:

$$B^A(i,j) = \begin{cases} 0, & ki \le j < k(i+1) \\ -\infty, & otherwise \end{cases} \quad (3)$$

By initializing the motion embedding from the subject ‘s neutral silent state rather than a learned or arbitrary representation, the decoder autoregressively predicts expression deformation offsets $\Delta\delta_k$ relative to each subject ‘s own articulatory baseline, enabling cross-identity adaptation without any subject-specific fine-tuning at inference time.

Training objectives

During the training phase, we adopt reconstruction loss and temporal consistency loss that jointly encourage accurate lip synchronization and temporally coherent mouth motion. We apply an ℓ_1 loss between the predicted driven keypoints x_k^{driven} and the ground-truth keypoints x_k^{gt} extracted from the target video frames:

$$\mathcal{L}_{rec} = \frac{1}{T}\sum_{t=1}^{T}\sum_{k\in\mathcal{S}} \parallel \mathbf{x}_{k,t}^{driven} - \mathbf{x}_{k,t}^{gt} \parallel_1 \tag{4}$$

To suppress inter-frame jitter and promote smooth articulatory transitions, we penalize abrupt changes in the predicted expression deformation offsets between consecutive frames:

$$\mathcal{L}_{temp} = \frac{1}{T-1}\sum_{t=2}^{T}\sum_{k\in\mathcal{S}} \parallel \Delta\delta_{k,t} - \Delta\delta_{k,t-1} \parallel_2^2 \tag{5}$$

The final training objective is a weighted combination of the two terms:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda\,\mathcal{L}_{temp} \tag{6}$$

where λ is a balancing hyperparameter that controls the trade-off between reconstruction accuracy and temporal smoothness, set to $\lambda = 0.1$ in our experiments.

Implicit Motion-Guided Latent Completion

As illustrated in Figure 1, the second stage adopts the pre-trained VAE model and the multimodal U-Net architecture from Latent Diffusion [25] as the backbone. Specifically, the mouth-masked source image and a reference image are encoded into the VAE latent space as the UNet input; the implicit expression deformation δ_k extracted from the source image via the LivePortrait motion extractor are injected into the UNet via cross-attention to guide completion of the masked mouth region. Building upon this foundation, we further introduce a Reference Spatial Cross-Attention (RSCA) module that injects reference appearance into the U-Net decoder via a query-based retrieval mechanism, enabling the network to selectively extract the lip and teeth texture it needs at each scale.

Latent Space Formulation

Unlike pixel-space methods such as Wav2Lip [3] and DINet [20], which conflate structural recovery and high-frequency texture synthesis within a single regression pass and consequently produce over-smoothed mouth regions, IMGTalk operates entirely within the latent space of a pre-trained, frozen VAE [25] model. Operating in the VAE latent space offers two key advantages over pixel-space: the significantly reduced spatial resolution lowers the computational cost of UNet-based completion, and the pre-trained VAE decoder provides a strong generative prior that facilitates high-fidelity reconstruction of fine-grained lip and teeth details from compact latent codes. The source image I_s is first masked in the mouth region to obtain $\tilde{I}_s = I_s \odot (1 - M)$, and then encoded by the frozen VAE encoder $\mathcal{E}$ to produce the masked source latent:

$$\tilde{\mathbf{z}}_s = \mathcal{E}(\tilde{I}_s) \in \mathbb{R}^{C \times h \times w}, \quad h = H/8, w = W/8 \tag{7}$$

The reference image I_r selected from the same identity with a similar head pose, is encoded in parallel as $z_r = \mathcal{E}(I_r) \in \mathbb{R}^{C \times h \times w}$, and routed exclusively through the RSCA module. The U-Net outputs a completed latent $\hat{z} \in \mathbb{R}^{C \times h \times w}$, which is decoded by the frozen VAE decoder $\mathcal{D}$ to produce the synthesized frame $\hat{I} = \mathcal{D}(\hat{z})$.

Implicit Motion condition

The implicit motion is extracted from the source image I_s using the frozen motion extractor of LivePortrait [14]. Although these offsets are implicit 3D vectors rather than explicit facial landmarks, LivePortrait demonstrates that they carry well-structured 3D facial information: the canonical keypoints encode subject identity, the rotation matrix encodes head pose, and the expression deformation $\delta_d \in \mathbb{R}^{K \times 3}$ represent pure expression displacement that is disentangled from both identity and head pose. Crucially, individual keypoints within δ_d have distinct semantic roles, for instance keypoints 6, 12, 14, 17, 19, and 20 are specifically associated with lip region dynamics, encoding the 3D structural deformation of the mouth during speech.

This structured semantics means that naively flattening δ_d into a single global vector would destroy the per-keypoint correspondence and lose the topological ordering that distinguishes expression movements. Instead, we preserve the sequential structure of the K keypoints by projecting each offset independently into a d_m -dimensional token via a shared linear layer, then summing with a learnable positional encoding $PE \in \mathbb{R}^{K \times d_m}$ that reinforces the spatial ordering among keypoints:

$$\mathbf{T} = Linear(\delta_k) \in \mathbb{R}^{K \times d_m}, \quad \mathbf{T}_{pe} = \mathbf{T} + \mathbf{PE} \tag{8}$$

The resulting sequence $T_{pe} \in \mathbb{R}^{K \times d_m}$ serves as Keys and Values in cross-attention layers across all U-Net layers, serving as the sole driving condition for mouth shape generation:

$$\mathbf{h}'_l = CrossAttn(\mathbf{Q} = \mathbf{h}_l, \mathbf{K} = \mathbf{T}_{pe}, \mathbf{V} = \mathbf{T}_{pe}) \tag{9}$$

Reference Spatial Cross-Attention

A common approach to injecting reference appearance in latent inpainting is to concatenate the reference latent z_r with the masked source latent along the channel dimension [4]. While simple to implement, this design has a fundamental limitation: reference information is mixed uniformly into the input regardless of what the network actually requires at each layer, forcing the U-Net to implicitly disentangle source context from reference appearance throughout all its computations.

We propose Reference Spatial Cross-Attention (RSCA), which injects reference appearance into every layer of the U-Net as Keys and Values, while the internal feature maps of each layer serve as Queries. This query-based formulation allows each spatial position at every scale to actively retrieve the most relevant appearance information from the reference on demand, rather than passively absorbing a fixed concatenated signal. The reference latent Z_r , already obtained from the frozen VAE encoder $\mathcal{E}$, is reused directly as the source of Keys and Values without any additional encoder. At each U-Net layer l , z_r is flattened into a sequence of hw spatial tokens and projected into the attention dimension via learned linear projections. Each layer maintains its own projection matrices, allowing different layers to learn to extract different aspects of the reference appearance. This produces a natural hierarchical specialization across the U-Net: shallow encoder layers tend to retrieve fine-grained texture and edge details; the bottleneck layer attends to global mouth structure and spatial relationships of dental arches; decoder layers progressively reconstruct appearance at increasing spatial resolution. This layer-wise division of attention is consistent with the general behavior of U-Net architectures. At every U-Net layer l , the projected reference tokens are injected via cross-attention in a residual formulation:

$$\mathbf{h}'_l = \mathbf{h}_l + \alpha_l \cdot CrossAttn(\mathbf{Q} = \mathbf{h}_l, \mathbf{K} = \mathbf{K}_r^l, \mathbf{V} = \mathbf{V}_r^l) \tag{10}$$

where α_l is a per-layer learnable scalar initialized to zero.

Training objectives

The implicit motion guided UNet is trained end-to-end with a composite objective that covers reconstruction fidelity, perceptual quality, motion reconstruction, and local texture realism. A pixel-level ℓ_1 loss is applied over the full image to supervise the generation:

$$\mathcal{L}_{rec} = \| \hat{I} - I_{gt} \|_1 \tag{11}$$

A perceptual loss over intermediate VGG-19 [27] feature layers mitigates the over-smoothing tendency of ℓ_1 -only training:

$$\mathcal{L}_{perc} = \sum_l \| \phi_l(\hat{I}) - \phi_l(I_{gt}) \|_2^2 \tag{12}$$

While $\mathcal{L}_{rec}$ and $\mathcal{L}_{perc}$ supervise appearance in the image space, they provide no direct geometric constraint on the correctness of the generated mouth shape. To enforce 3D structural consistency between the synthesized mouth and the driving motion, and to prevent identity drift during generation, we extract implicit motion from the generated frame $\hat{I}$ and compute an ℓ_1 distance against those extracted from the source image:

$$\mathcal{L}_{motion} = \| \hat{\delta}_k - \delta_k \|_1 + \lambda_c \| \hat{\mathbf{x}}_c - \mathbf{x}_c \|_1 \tag{13}$$

where $\hat{\delta}_k$ term penalizes deviations in the implicit 3D mouth structure, and $\hat{x}_c$ term constrains the generated identity to remain consistent with the source. A patch-based discriminator $\mathcal{D}_{patch}$ further promotes local texture realism and suppresses boundary artifacts at the mask edges:

$$\mathcal{L}_{adv} = -\mathbb{E}[\log \mathcal{D}_{patch}(\hat{I})] \tag{14}$$

The total training objective is a weighted linear combination of the four terms:

$$\mathcal{L}_{total} = \lambda_{rec}\,\mathcal{L}_{rec} + \lambda_{perc}\,\mathcal{L}_{perc} + \lambda_{motion}\,\mathcal{L}_{motion} + \lambda_{adv}\,\mathcal{L}_{adv} \tag{15}$$

EXPERIMENTS

Experimental Settings

Datasets

We conduct experiments on two publicly available audio-visual datasets. The HDTF dataset [28] is a high-quality collection of talking-face videos recorded at 720P–1080P resolution, sourced from YouTube. It covers 348 videos and provides approximately 16 hours of footage. The TFHP dataset [29] contains 1,052 videos of 588 subjects totaling 26.5 hours, spanning a diverse range of speaking contexts, including lectures, interviews, and news broadcasts. For both datasets, videos are processed at 25 FPS, the accompanying audio tracks are resampled to 16 kHz, and face regions are cropped and resized to 256x256 pixels. For each frame, the implicit motion representation is pre-extracted and cached to avoid redundant computation during training. For each video, a single reference frame is selected as the neutral state: we identify the frame whose head pose is closest to a frontal orientation and whose mouth is in a closed resting position, and its extracted motion serves as the silent-state embedding used in Stage 1. We follow the standard speaker-level split for each dataset, partitioning subjects into disjoint training, validation, and test sets.

Baselines

We compare IMGTalk against five representative talking-face generation methods: (1) Wav2Lip [3]: pioneering GAN-based method employing a pre-trained lip-sync discriminator to enforce audio-visual correspondence directly in pixel space; (2) IP-LAP [11]: two-stage framework combining Transformer-based landmark prediction with multi-reference feature alignment for identity-preserving lip synchronization; (3) DINet [20]: deformation-based network that spatially warps reference feature maps guided by audio features; (4) VideoReTalking [21]: three-stage pipeline that sequentially normalizes facial expression, performs lip synchronization, and enhances face quality for high-fidelity video dubbing; (5) MuseTalk [4]: real-time latent space inpainting method that concatenates masked source and reference latents as UNet input conditioned on audio features.

Evaluation metrics

We adopt five complementary metrics for quantitative evaluation. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) measure pixel-level reconstruction fidelity and perceptual structural integrity against ground-truth frames, respectively. Fréchet Inception Distance (FID) [30] measures the distributional similarity between generated and real frame patches, providing a perceptual photorealism as-

sessment. Landmark Distance (LMD) [1] quantifies the geometric accuracy of lip articulation by computing the mean Euclidean distance between predicted and ground-truth facial landmarks in the mouth region, providing a direct measure of lip-sync correctness. Audio-Visual Confidence (AVConf) evaluates audio-visual synchronisation naturalness via SyncNet 's cross-modal embedding similarity, assessing dynamic lip-sync quality beyond static geometric alignment. Together, these metrics collectively cover pixel-level fidelity, distributional photo-realism, geometric lip-sync accuracy, and audio-visual synchronisation, providing a comprehensive evaluation of both reconstruction quality and generative perceptual quality.

Quantitative Evaluation

The quantitative experimental results are summarized in Table 1. Our proposed IMGTalk achieves the best performance across all four evaluation metrics on both the HDTF and TFHP datasets, demonstrating its superior capability in visual quality, geometric fidelity, and audio-visual synchronization. Regarding geometric accuracy, our method achieves the lowest LMD, which indicates that the implicit motion guidance provides more precise constraints for mouth shape than traditional landmarks or raw audio features. Notably, for the AVConf metric, IMGTalk also achieves the highest scores, slightly surpassing the state-of-the-art methods Wav2Lip and MuseTalk. It is worth emphasizing that both Wav2Lip and MuseTalk incorporate an explicit SyncLoss during their training phases to supervise audio-visual alignment, derived from a pre-trained SyncNet. Since the AVConf metric is also computed using a SyncNet-based architecture, these baselines naturally hold a structural advantage for this particular metric. Despite this disadvantage, IMGTalk still achieves the highest AVConf scores on both datasets, demonstrating that geometrically grounded implicit motion conditioning provides a stronger and more generalizable audio-visual alignment signal than direct SyncLoss supervision.

Table 1. Quantitative comparisons with baselines on the HDTF and TFHP datasets

Method	HDTF SSIM↑	HDTF PSNR↑	HDTF FID↓	HDTF LMD↓	HDTF AV-Conf↑	TFHP SSIM↑	TFHP PSNR↑	TFHP FID↓	TFHP LMD↓	TFHP AV-Conf↑
Ground Truth	1	N/A	N/A	0	8.839	1	N/A	N/A	0	8.516
Wav2Lip	0.718	26.352	12.763	4.285	7.637	0.745	25.026	12.036	4.338	7.524
VideoRetalking	0.862	29.524	11.652	3.735	6.847	0.983	29.123	10.977	3.219	6.621
IPLAP	0.798	24.956	12.167	3.642	5.865	0.821	25.162	11.658	4.987	5.252
DINet	0.942	30.008	7.243	2.190	7.275	0.928	30.687	7.039	2.268	7.556
MuseTalk	0.958	33.615	6.792	2.013	7.612	0.976	33.132	6.541	2.008	7.580
ours	0.969	34.259	6.135	1.785	7.697	0.984	34.167	6.092	1.961	7.613

Qualitative Comparison

Representative qualitative results are shown in Figure 2, where we compare IMGTalk against all baselines on test samples drawn from the HDTF and TFHP datasets. As illustrated in Figure 2, Wav2Lip tends to produce blurry and over-smoothed mouth regions, losing fine-grained lip and teeth texture due to its pixel-space regression strategy. VideoReTalking introduces jagged boundary artifacts around the lip contour and over-smooths the perioral region, occasionally causing unnatural facial stiffness. DINet, while deformation-based, induces noticeable identity drift in the generated results, as its audio-driven warping field lacks explicit geometric constraints on mouth shape. IP-LAP maintains identity relatively well owing to its landmark guidance, but pro-duces inconsistent lip movements and struggles to faithfully reproduce extreme mouth apertures, resulting in blurred inter-dental regions. MuseTalk achieves a better balance among lip movement consistency, identity preservation, and inference efficiency by operating in the VAE latent space, yet its audio-only conditioning causes it to regress toward identity-averaged mouth dynamics, leading to under-animated lip openings and inconsistent teeth visibility across frames.

In contrast, IMGTalk generates sharper lip and teeth textures, more precise mouth articulation, and cleaner mask boundaries across all test cases. The implicit motion conditioning provides geometrically grounded constraints on mouth shape at each frame, preventing the identity-averaged collapse seen in MuseTalk and the identity drift observed in DINet. The RSCA module further enables selective retrieval of fine-grained appearance details from the reference frame, contributing to faithful reconstruction of teeth structure and lip texture.

Figure 2. Qualitative comparison with baselines on HDTF test samples

Each row shows results from the same driving audio. Readers are encouraged to zoom in on the mouth region to inspect lip and teeth texture fidelity. Our method produces the sharpest inter-dental structure, most precise lip articulation, and cleanest mask boundaries across all subjects.

Ablation Study

To evaluate the effectiveness of the proposed IMGTalk, we conduct comprehensive ablation experiments on the HDTF dataset, with quantitative results summarized in Table 2. We first investigate the impact of our Implicit Motion (IM) representation by replacing it with traditional explicit priors, specifically 3DMM coefficients (w/ 3DMM) and facial landmarks (w/ landmark). The observed increase in LMD and the loss of subtle emotional expressions confirm that traditional parametric models often suffer from representation gaps and jitter, whereas our IM provides more fluid and precise geometric guidance for mouth synchronization. Furthermore, we verify the superiority of the Reference Spatial Cross-Attention (RSCA) by replacing it with a standard channel-wise concatenation (w/o RSCA). The drop in PSNR and SSIM indicates that while concatenation provides global context, RSCA allows the network to adaptively retrieve fine-grained textures from the reference frame. Finally, removing the motion reconstruction loss (w/o $\mathcal{L}_{motion}$) leads to 3D structural incon-

sistencies during rapid speech, demonstrating its indispensable role in maintaining the geometric integrity of the latent completion process. In conclusion, the synergy of these components enables IMGTalk to achieve superior generative quality and audio-visual alignment.

Table 2. Ablation study of implicit motion, RSCA components and the motion reconstruction loss on HDTF

Method	SSIM↑	PSNR↑	FID↓	LMD↓	AVConf↑
w/ landmark	0.948	33.871	6.545	1.976	7.058
w/ 3DMM	0.953	33.529	6.428	2.116	7.160
w/o RSCA	0.914	31.246	6.736	1.826	7.568
w/o $\mathcal{L}_{motion}$	0.950	34.125	6.207	2.108	6.922
Ours (full)	0.969	34.259	6.135	1.785	7.697

CONCLUSION

In this paper, we presented IMGTalk, a novel two-stage implicit motion-guided latent completion framework for audio-driven talking head synthesis. By decoupling the task into a generalizable Audio-to-Implicit Motion Encoding stage and a high-fidelity Latent Space Completion stage, our method effectively bridges the modality gap between acoustic input and photorealistic visual synthesis. Specifically, the proposed Implicit Motion (IM) representation provides stronger geometric constraints than traditional landmarks or 3DMM parameters, ensuring precise lip synchronization without the need for an explicit SyncLoss. Furthermore, the Reference Spatial Cross-Attention (RSCA) module enables adaptive retrieval of fine-grained appearance information, significantly improving identity preservation and teeth/lip texture reconstruction. Quantitative and qualitative evaluations on HDTF and TFHP datasets demonstrate that IMGTalk outperforms state-of-the-art methods in all key metrics, achieving superior visual quality and audio-visual alignment.

Limitations and Future Work

Despite promising results, there remain several avenues for future exploration. Currently, IMGTalk primarily focuses on mouth region synthesis while keeping the head pose consistent with the source image. Future research could integrate a dedicated head pose generation module to enable more natural and expressive head movements. Additionally, we plan to extend the framework to handle more extreme emotional expressions and diverse lighting conditions to further enhance the robustness of zero-shot generation in the wild.

Ethical Considerations

IMGTalk enables the synthesis of realistic talking-head videos from a single reference image or video clip driven solely by an audio input, a capability that carries inherent dual-use risks. The most immediate concern is identity misappropriation: the framework could be exploited to fabricate videos of real individuals speaking content they never uttered, facilitating impersonation, misinformation, and reputational harm. We strongly advocate that any real-world deployment be subject to explicit informed consent from the individuals whose likeness is used, and that all synthetic content be clearly attributed as machine-generated to prevent deceptive dissemination.

Author Contributions

Conceptualization – Xiang Li; methodology – Xiang Li; formal analysis – Xiaoli Huang; investigation – Xiang Li and Xiaoli Huang; resources – Xiang Li; writing-original draft preparation – Xiang Li and Xiaoli Huang; writing-review and editing – Xiang Li, Xiaonan Luo and Xiaoli Huang; visualization – Xiaoli Huang; supervision – Xiaonan Luo and Xiaoli Huang. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

This work was supported by the Guangxi Science and Technology Major Program under GuikeAA24263013.

Acknowledgements

Not applicable.

REFERENCES

[1] Chen, L., Li, Z., Maddox, R.K., Duan, Z., Xu, C.: Lip movements generation at a glance. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 520–535 (2018). http://doi.org/10.1007/978-3-030-01234-2_32

[2] Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7832–7841 (2019). http://doi.org/10.1109/cvpr.2019.00802

[3] Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.V.: 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. MM '20, pp. 484–492. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3394171.3413532

[4] Zhag, Y., Zhong, Z., Liu, M., Chen, Z., Wu, B., Zeng, Y., Zhan, C., He, Y., Huang, J., Zhou, W.: Musetalk: Real-time high-fidelity video dubbing via spatio- temporal sampling. arxiv (2025). https://arxiv.org/pdf/2410.10122

[5] Tian, L.; Wang, Q.; Zhang, B.; Bo, L. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 244–260. http://doi.org/10.1007/978-3-031-73010-8_15

[6] Stypułkowski, M., Vougioukas, K., He, S., Zięba, M., Petridis, S., Pantic, M.: Diffused heads: Diffusion models beat gans on talking-face generation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5091–5100 (2024). https://arxiv.org/pdf/2410.10122

[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). http://arxiv.org/pdf/2412.00733

[8] Huang, R., Lai, P., Qin, Y., Li, G.: Parametric implicit face representation for audio-driven facial reenactment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12759–12768 (2023). https://openaccess.thecvf.com/content/CVPR2023/html/Huang_Parametric_Implicit_Face_Representation_for_Audio-Driven_Facial_Reenactment_CVPR_2023_paper.html

[9] Huang, X., Li, X., Tan, S., Chen, W., Li, G.: Gaptalk: precision-controlled 3d gaussian rendering for personalized talking-head synthesis. The Visual Computer 41(15), 12953–12966 (2025). https://link.springer.com/article/10.1007/s00371-025-04195-y

[10] Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M.: Neural voice puppetry: Audio-driven facial reenactment. In: European Conference on Computer Vision, pp. 716–731 (2020). http://arxiv.org/pdf/1912.05566

[11] 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9729–9738 (2023). https://openaccess.thecvf.com/content/CVPR2023/papers/Zhong_Identity-Preserving_Talking_Face_Generation_With_Landmark_and_Appearance_Priors_CVPR_2023_paper.pdf

[12] Zhang, C., Zhao, Y., Huang, Y., Zeng, M., Ni, S., Madhukar Budagavi, et al.: Facial: Synthesizing dynamic talking face with implicit attribute learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3867–3876 (2021). http://doi.org/10.1109/iccv48922.2021.00384

[13] 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 (2024). Springer. https://link.springer.com/chapter/10.1007/978-3-031-72658-3_23

[14] 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). https://arxiv.org/pdf/2407.03168?

[15] Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D.B., Genova, K., Jin, Z., Theobalt, C., Agrawala, M.: Text-based editing of talking-head video. acm Transactions on Graphics (tog) 38(4), 1–14 (2019). http://doi.org/10.1145/3306346.3323028

[16] Guo, Y., Chen, K., Liang, S., Liu, Y.-J., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5784–5794 (2021). http://doi.org/10.1109/iccv48922.2021.00573

[17] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM 65(1), 99–106 (2021). https://dl.acm.org/doi/pdf/10.1145/3503250

[18] Tang, J., Wang, K., Zhou, H., Chen, X., He, D., Hu, T., Liu, J., Liu, Z., Zeng, G., Wang, J.: Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. International Journal of Computer Vision, 1–12 (2025). http://link.springer.com/10.1007/s11263-025-02481-9

[19] Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. 42(4), 139–1 (2023). https://sgvr.kaist.ac.kr/~sungeui/ICG_F23/Students/[CS482]%203D%20Gaussian%20Splatting%20for%20Real-Time%20Radiance%20Field%20Rendering.pdf

[20] Zhang, Z., Hu, Z., Deng, W., Fan, C., Lv, T., Ding, Y.: Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 3543–3551 (2023). http://doi.org/10.1609/aaai.v37i3.25464

[21] Cheng, K., Cun, X., Zhang, Y., Xia, M., Yin, F., Zhu, M., Wang, X., Wang, J., Wang, N.: Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In: SIGGRAPH Asia 2022 Conference Papers, pp. 1–9 (2022). https://dl.acm.org/doi/pdf/10.1145/3550469.3555399

[22] Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. IEEE Transactions on pattern analysis and machine intelligence 25(9), 1063–1074 (2003). https://ieeexplore.ieee.org/abstract/document/1227983/

[23] Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Transactions on Graphics 36(6), 194–1 (2017). http://doi.org/10.1145/3130800.3130813

[24] Wang, T.-C., Mallya, A., Liu, M.-Y.: 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. 10039–10049 (2021). http://doi.org/10.1109/cvpr46437.2021.00991

[25] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022). https://arxiv.org/abs/2112.10752

[26] Hsu, W.-N., Bolte, B., Tsai, Y.-H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM transactions on audio, speech, and language processing 29, 3451–3460 (2021). http://doi.org/10.1109/taslp.2021.3122291

[27] Simonyan, Karen , and A. Zisserman . “Very Deep Convolutional Networks for Large-Scale Image Recognition.” Computer Science (2014). http://arxiv.org/pdf/1409.1556

[28] Zhang, Z., Li, L., Ding, Y., Fan, C.: 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). http://doi.org/10.1109/cvpr46437.2021.00366

[29] Sun, Z., Lv, T., Ye, S., Lin, M., Sheng, J., Wen, Y.-H., Yu, M., Liu, Y.-j.: Diff-posetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. ACM Transactions on Graphics (TOG) 43(4), 1–9 (2024). https://dl.acm.org/doi/pdf/10.1145/3658221

[30] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 30 (2017). http://arxiv.org/pdf/1706.08500