# MeanTalker: Efficient and Expressive Speech-Driven 3D Facial Animation via Geometric-Aware Mean Flow

Zhongyuan Zhao[1,2], Zhihao Li[1,2], Qing Li[2], Leidong Fan[2], and Kanglin Liu[2⋆]

[1] School of Electronic and Computer Engineering, Peking University, Shenzhen, China

[2] Pengcheng Laboratory, Shenzhen, China

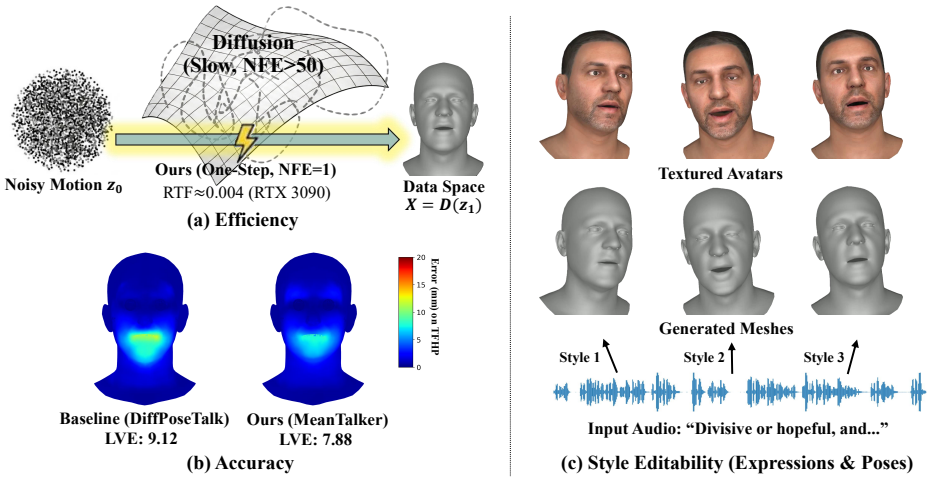


**Fig. 1:** We present MeanTalker, a novel framework for efficient and expressive speech-driven 3D facial animation. By bypassing slow iterative diffusion, it achieves one-step inference with an RTF of $\approx 0.004$ (a). Compared to the diffusion baseline, MeanTalker delivers superior lip-sync accuracy (b), while enabling geometrically consistent and stylistically diverse control over expressions and head poses from the same audio (c).

**Abstract.** Speech-driven 3D facial animation aims to synthesize natural and synchronized facial motions from arbitrary speech audio. Deterministic models struggle with the complex many-to-many mapping, thus yielding over-smoothed motions and lacking expression nuances. In comparison, diffusion-based approaches excel in capturing expressive distributions but suffer from significant latency due to iterative sampling. To this end, we propose MeanTalker, a novel one-step speech-to-motion generative framework built upon Mean Flow. It utilizes a Dual-Time Denoising Transformer (DTDT) for training stability and Geometric-Aware Trajectory Learning (GATL) for precise manifold alignment. Specifically,

⋆ Corresponding author: max.liu.426@gmail.com.

DTDT introduces an interleaved dual-time encoding to support a progressive curriculum from modeling instantaneous to average velocities, effectively circumventing the instability of direct mean flow training. Furthermore, GATL rectifies the generation trajectory through synergistic dual-space supervision. By applying explicit vertex constraints and projecting latent velocity errors onto the surface manifold, it strictly regularizes the flow direction to guarantee precisely synchronized one-step synthesis. Extensive experiments demonstrate that MeanTalker establishes state-of-the-art lip-sync accuracy while accelerating inference by over 200× compared to iterative diffusion models. Achieving an exceptional Real-Time Factor (RTF) of 0.004, our method effectively bridges the gap between precise motion generation and real-time deployment.

**Keywords:** Speech-driven · One-step Speech-to-motion · Mean Flow

## 1 Introduction

Speech-driven 3D facial animation constitutes a pivotal technique in computer graphics, facilitating the automated generation of digital avatars for virtual reality, gaming, and animation production [18,29]. The fundamental objective of this task is to synthesize synchronized and natural time-varying 3D mesh sequences, principally driven by input acoustic signals alongside stylistic control factors. Distinct from 2D video generation tasks which focus on texture rendering, 3D facial animation necessitates the precise prediction of vertex displacements [4, 5] or blendshape coefficients [21, 33] to animate a predefined topological template. Despite significant progress, achieving a synergy between expressive motion dynamics and real-time inference latency remains a formidable challenge.

To capture the diverse distribution of facial motions, probabilistic approaches based on diffusion models have recently been adopted [2, 15, 24, 26]. These methods model the conditional distribution of facial dynamics, successfully generating diverse and expressive animations that surpass regression baselines in terms of motion variance. However, this performance gain comes at the cost of significant latency. The iterative nature of the diffusion sampling process, which typically necessitates tens or hundreds of sequential denoising steps to approximate the data distribution, introduces a computational bottleneck that precludes their deployment in real-time interactive scenarios.

To address this challenge, we propose MeanTalker, a novel one-step speech-to-motion framework built upon Mean Flow [7] that preserves the expressive precision of generative models while enabling real-time inference, as illustrated in Fig. 1. Instead of relying on the slow, curved trajectories of standard diffusion processes, our approach learns a direct transport map that connects the noise distribution to the facial motion data distribution via a straight-line trajectory. To implement this effectively, we design a Dual-Time Denoising Transformer (DTDT) featuring an interleaved dual-time encoding mechanism. This architecture supports a progressive curriculum that smoothly transitions from modeling instantaneous velocities [16, 17] to average velocities [7], establishing

a robust generative baseline. Furthermore, to resolve the manifold deviation and geometric ambiguity inherent in purely latent-space generation, we introduce a Geometric-Aware Trajectory Learning (GATL) strategy. GATL rectifies the generation trajectory through synergistic dual-space supervision: it applies explicit vertex constraints on implicitly recovered endpoints to ensure physical validity, and projects latent velocity errors onto the surface manifold to strictly regularize the flow direction using geometric semantics. By anchoring the latent trajectory shaping within the physical domain, MeanTalker bridges the gap between latent dynamics and explicit topologies, guaranteeing precisely synchronized one-step synthesis. Extensive evaluations on standard benchmarks demonstrate that MeanTalker establishes new state-of-the-art performance, surpassing recent diffusion-based and autoregressive methods in lip-sync accuracy while accelerating inference by orders of magnitude.

In summary, our main contributions are as follows:

- We propose MeanTalker, an efficient framework for speech-driven 3D facial animation that overcomes the latency limitations of diffusion models through direct trajectory modeling based on Mean Flow.
- We develop a Dual-Time Denoising Transformer (DTDT) featuring a dual-time encoding mechanism that enables progressive modeling from instantaneous to average velocity fields for robust 3D facial motion generation.
- We propose Geometric-Aware Trajectory Learning (GATL), which synergistically integrates implicit manifold projection and semantic-aware trajectory rectification to ensure geometric validity and path straightness.
- Extensive experiments demonstrate that MeanTalker achieves state-of-the-art results on standard benchmarks, significantly outperforming existing methods in both lip-sync accuracy and inference speed.

## 2 Related Work

### 2.1 Speech-Driven 3D Facial Animation

Research in speech-driven 3D facial animation has transitioned from early procedural systems relying on static phoneme-viseme mapping rules [28] to data-driven deep learning frameworks [10,35] that learn complex audio-motion correlations. While procedural methods offer explicit control, they lack natural motion nuances. Consequently, contemporary approaches [3,26,32] predominantly focus on learning the mapping from acoustic signals to facial geometry, with different modeling choices spanning deterministic regression, probabilistic generation, and efficient generative modeling for real-time deployment.

### 2.2 Deterministic Regression Approaches

Deterministic regression methods formulate facial animation as a direct mapping task, typically categorized into vertex-based and parameter-based approaches.

In the vertex-based domain, VOCA [4] pioneered the use of temporal convolutional networks to regress vertex offsets from DeepSpeech [9] features, conditioned on subject identity. Building upon this, FaceFormer [5] introduced a Transformer [30] architecture with periodic positional encoding and biased attention mechanisms to better capture long-term audio context and ensure lip closure. To address the issue of over-smoothing, CodeTalker [32] cast the problem as a discrete code query task within a learned finite motion space, reducing the uncertainty of the cross-modal mapping. Parameter-based approaches like SadTalker [33] and EmoTalk [21] predict coefficients for 3D Morphable Models (3DMM) [1], effectively disentangling head pose, expression, and emotion through specialized encoders. VividTalk [25] further adopts a hybrid strategy, utilizing blendshapes for coarse global motion and vertex offsets for fine-grained lip details. Despite much progress, deterministic methods still face the inherent ambiguity of cross-modal generation, since a single speech segment can correspond to multiple valid facial expressions and styles. As a result, deterministic optimization tends to drive the model toward the conditional expectation, leading to reduced expressive diversity and vividness, where generated motions may appear stiff or overly stabilized compared with real speakers.

### 2.3 Probabilistic Generative Approaches

To mitigate the over-smoothing limitations of deterministic mapping and capture rich facial dynamics, recent research has shifted towards probabilistic generative models that learn the full conditional distribution of facial motions. Among these methods, Denoising Diffusion Probabilistic Models (DDPMs) [11] have emerged as a powerful continuous-space paradigm. FaceDiffuser [24] pioneered this direction by employing a GRU-based decoder to iteratively denoise motion sequences from Gaussian noise, successfully generating expressive animations. Extending this capability, DiffPoseTalk [26] incorporated a style encoder and classifier-free guidance [12] to jointly generate stylized expressions and head poses of the FLAME [13] model, allowing for reference-based style control. More recently, MSMD [20] further explored example-based style conditioning by combining a VAE style encoder with a latent diffusion model and a style basis that guides generation from reference clips. Similarly, GLDiTalker [15] combined diffusion with a quantized latent space to alleviate audio-mesh modality misalignment. A related but distinct branch studies emotion-controllable generation. ProbTalk3D [31] adopts a two-stage VQ-VAE framework to generate non-deterministic facial motions conditioned on speech, identity, emotion category, and emotion intensity. This line of work focuses on explicit emotion control using emotion-rich datasets, whereas our approach considers more general speaking styles without relying on explicit emotion annotations. While recent attempts like DiffusionTalker [2] explore model distillation to accelerate sampling, such methods often suffer from quality degradation. Overall, diffusion-based approaches still rely on iterative denoising, and balancing expressive motion generation with real-time inference remains challenging.

### 2.4 Efficient Generative Modeling

Efficiency is also critical for deploying speech-driven 3D facial animation in interactive scenarios. For this task, ARTalk [3] achieves real-time generation with an autoregressive model over discrete motion codes. Tiny-V2F [8] explores a complementary efficiency-oriented direction by distilling large speech-driven animation models into compact models for resource-constrained inference. A similar efficiency concern has also been studied in diffusion and flow based generative models, where the main goal is to reduce the cost of iterative sampling. Consistency models [23] learn mappings from noisy states to clean samples under a consistency constraint. Shortcut models [6] condition the model on the desired step size, while inductive moment matching [34] trains efficient generators by matching intermediate marginal distributions. Closely related to our formulation, Mean Flow [7] models average velocity rather than instantaneous velocity for direct one-step generation. Motivated by this line of work, MeanTalker adapts mean-flow generation to speech-driven 3D facial animation, where framewise audio alignment, lip synchronization, and FLAME manifold validity must be jointly preserved.

Overall, existing methods reveal two complementary requirements for speech-driven 3D facial animation: expressive motion modeling and efficient real-time synthesis. Regression-based methods are efficient but tend to suppress the diversity of valid facial motions, while diffusion-based methods improve expressiveness at the cost of iterative sampling. Recent efficient generative models reduce this sampling cost, but directly applying them to 3D facial animation still leaves audio to lip alignment and geometric validity insufficiently constrained. MeanTalker is designed to bridge this gap by combining Mean Flow based one-step generation with a Dual-Time Denoising Transformer (DTDT) and Geometric-Aware Trajectory Learning (GATL), enabling expressive and synchronized FLAME motion in a single forward pass.

## 3 Method

This section details the architecture and learning objectives of MeanTalker. As illustrated in Fig. 2, our framework is structurally composed of two core synergistic components that collaborate for highly efficient one-step speech-driven facial animation. We begin by establishing the theoretical foundations of Conditional Flow Matching and the Mean Flow Identity in Sec. 3.1. Subsequently, we detail the **Dual-Time Denoising Transformer** (Part 1, Sec. 3.2), which integrates multi-modal acoustic-style conditions with a novel dual-time encoding mechanism to predict the generative velocity fields. Building upon this, we elaborate on the **Geometric-Aware Trajectory Learning** module (Part 2, Sec. 3.3), which imposes exact manifold projection and physics-guided Mean Flow supervision to effectively straighten the generation path. Finally, we summarize the optimization objectives and the training strategy in Sec. 3.4.

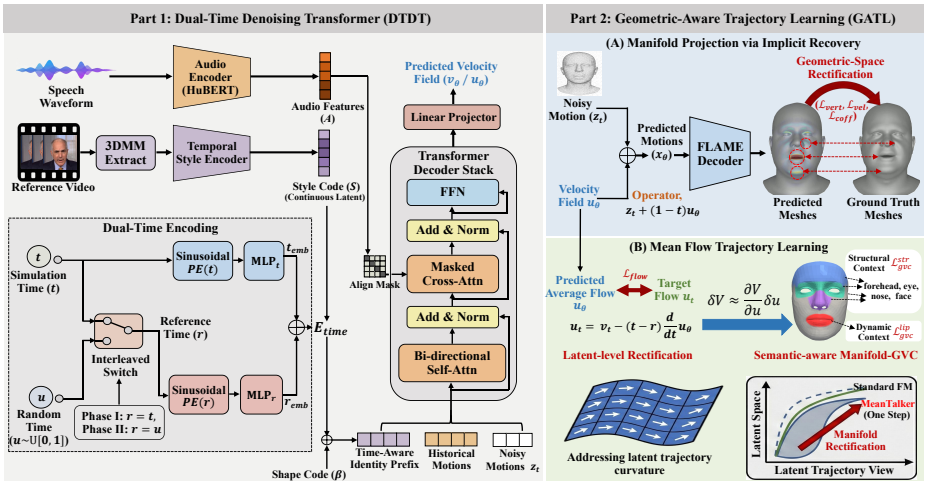


**Fig. 2: Overview of MeanTalker framework. Part 1: DTDT** predicts generative velocity fields ($v_\theta/u_\theta$) via a Dual-Time Encoding module. An interleaved switch between simulation time $t$ and reference time $r$ supports both instantaneous and average velocities modeling. **Part 2: GATL** ensures robust one-step inference through: (A) Manifold Projection that recovers endpoint $x_\theta$ for geometric alignment; and (B) Mean Flow Trajectory Learning that aligns $u_\theta$ with JVP-rectified targets $u_t$, incorporating the Semantic-aware Manifold-GVC strategy ($\mathcal{L}_{gvc}^{lip}, \mathcal{L}_{gvc}^{str}$) to project latent trajectory rectification onto the interpretable surface manifold for region-specific shaping.

### 3.1 Preliminaries

**Conditional Flow Matching (CFM).** CFM [16] provides a simulation-free framework for training Continuous Normalizing Flows to generate data from a noise distribution. Let $x_1 \sim q(x_1|c)$ denote the target facial motion distribution conditioned on the acoustic and style context $c$, and let $x_0 \sim p(x_0)$ be samples from a standard Gaussian prior $\mathcal{N}(0, I)$. To construct a probability density path between $p_0$ and $q$, we adopt the optimal transport formulation, which defines a linear interpolation trajectory: $z_t = (1-t)x_0 + tx_1$, for time $t \in [0, 1]$. The probability density path is governed by an ordinary differential equation (ODE) $dz_t/dt = v_t(z_t)$, where $v_t$ is a time-dependent vector field. The target conditional vector field $v_t(z_t|x_1)$ that generates this straight-line path is constant and points directly from noise to data:

$$v_t(z_t|x_1) = \frac{d}{dt} z_t = x_1 - x_0 \tag{1}$$

A neural network $v_\theta$ is trained to approximate this instantaneous velocity field by minimizing the expected mean squared error:

$$\mathcal{L}_{cfm}(\theta) = \mathbb{E}_{t,x_0,x_1}\left[\|v_\theta(z_t, t, c) - (x_1 - x_0)\|^2\right] \tag{2}$$

During inference, generating motions requires solving the ODE using iterative numerical solvers. While standard CFM successfully models local instantaneous

velocities, this reliance on numerical integration necessitates multiple function evaluations (NFEs), imposing severe latency bottlenecks for real-time applications.

**Mean Flow Identity.** To circumvent the iterative ODE solving, Mean Flow [7] shifts the modeling target from the instantaneous velocity $v(z_t, t)$ to an average velocity $u(z_t, r, t)$. The average velocity is defined as the integral of the instantaneous velocity over the time interval from a reference time $r$ to the current time $t$, divided by the time interval:

$$u(z_t, r, t) \triangleq \frac{1}{t-r} \int_r^t v(z_\tau, \tau) d\tau \tag{3}$$

By taking the total derivative with respect to $t$, this definition elegantly yields the Mean Flow Identity:

$$u(z_t, r, t) = v(z_t, t) - (t-r) \frac{d}{dt} u(z_t, r, t) \tag{4}$$

This identity demonstrates a fundamental, simulation-free relationship between the average and instantaneous velocities. By constructing a network $u_\theta$ to satisfy Eq. (4), the model learns to directly predict the straight-line displacement towards the target. At inference, setting $r = 1$ allows for exact one-step generation via $x_1 = z_t + (1-t)u_\theta(z_t, 1, t)$, completely eliminating the need for iterative integration.

### 3.2 Dual-Time Denoising Transformer

The Dual-Time Denoising Transformer (DTDT) serves as the core dynamics engine of MeanTalker, explicitly designed to translate acoustic features into expressive 3D facial motions. For highly complex cross-modal generation tasks, directly learning the average velocity field from scratch often suffers from severe training instability [14]. To circumvent this bottleneck, our DTDT introduces a dual-time encoding mechanism. This design establishes a progressive curriculum learning paradigm, granting the network the flexibility to transition smoothly from modeling the instantaneous dynamics of audio-lip synchronization to predicting the global average velocity for one-step facial motion generation.

**Dual-Time Encoding and Architecture.** Defining an average velocity for motion generation necessitates two temporal anchors: the current simulation time $t$ and a target reference time $r$. To explicitly model this temporal span, we sample $t \sim \mathcal{U}[0, 1]$ and an independent random time $u \sim \mathcal{U}[0, 1]$ during training. To operationalize our progressive curriculum (detailed in Sec. 3.4), we first pretrain the model exclusively to capture fundamental cross-modal correlations, and subsequently fine-tune it by alternating between predicting instantaneous and

average velocities. During this process, an interleaved switch dynamically routes the reference time $r$:

$$r = \begin{cases} t, & \text{if Phase I (Instantaneous Velocity)} \\ u, & \text{if Phase II (Average Velocity)} \end{cases} \tag{5}$$

Phase I forces the network to degenerate into predicting the instantaneous velocity $v_\theta(z_t, t)$, establishing a robust generative baseline for speech-driven facial movements. Phase II exposes the network to arbitrary time horizons to predict the global average velocity $u_\theta(z_t, r, t)$ for straight-line trajectory generation. The dual scalars are expanded via Sinusoidal Positional Encoding (PE) and projected through dedicated Multi-Layer Perceptrons ($\text{MLP}_t$ and $\text{MLP}_r$) to form the composite time embedding $E_{time} = \text{MLP}_t(\text{PE}(t)) + \text{MLP}_r(\text{PE}(r))$.

To ground the facial animation in a specific identity, we extract a static shape coefficient $\beta$ from the reference video. Concurrently, we employ a pre-trained and frozen style encoder [26] to extract a temporal style code $S$ from the reference motion sequence. These features are concatenated to form a comprehensive identity code $I = [\beta; S]$. Subsequently, $E_{time}$ is integrated into $I$ via element-wise addition to construct a Time-Aware Identity Prefix $C_{prefix} = I + E_{time}$. The Transformer Decoder Stack processes the composite sequence $[C_{prefix}, X_{hist}, z_t]$ using Bi-directional Self-Attention for global context modeling. To ensure strict phoneme-to-lip synchronization, a Masked Cross-Attention mechanism is applied with a diagonal alignment mask, forcing each motion frame within $z_t$ to attend to its corresponding local audio window in the acoustic feature $A$. Ultimately, the DTDT predicts the velocity field ($v_\theta$ or $u_\theta$) for subsequent trajectory straightening in the GATL module.

### 3.3 Geometric-Aware Trajectory Learning

While the DTDT predicts the generative field, standard flow matching objectives often suffer from curvature error during one-step inference, leading to off-manifold artifacts. To address this, we propose Geometric-Aware Trajectory Learning (GATL), which rectifies the generation path via synergistic latent and manifold level constraints.

**Manifold Projection via Implicit Recovery.** To impose geometric constraints without iterative sampling, we leverage the predictive power of the average velocity $u_\theta$ to perform implicit motion recovery. As illustrated in Fig. 2(A), for any state $z_t$ sampled at time $t$, we analytically anticipate the clean motion endpoint $x_\theta$:

$$x_\theta = z_t + (1 - t) \cdot u_\theta(z_t, 1, t) \tag{6}$$

By mapping the recovered parameters $x_\theta$ to the 3D vertex space via the FLAME decoder, we can supervise the generation at the physical level. The geometric objective $\mathcal{L}_{geo}$ is defined as:

$$\mathcal{L}_{geo} = \lambda_{vert}\mathcal{L}_{vert} + \lambda_{vel}\mathcal{L}_{vel} + \lambda_{coef}\mathcal{L}_{coef} \tag{7}$$

where $\mathcal{L}_{vert}$ ensures precise lip synchronization, $\mathcal{L}_{vel}$ eliminates temporal jitters, and $\mathcal{L}_{coef}$ maintains boundary continuity via structured coefficient constraints.

**Mean Flow Trajectory Learning.** To eliminate trajectory curvature, we supervise the predicted average velocity $u_\theta$ against a JVP-rectified target $u_t$:

$$u_t = (x_1 - x_0) - (t - r)\frac{d}{dt}u_\theta(z_t, r, t) \tag{8}$$

where the total derivative is computed via the Jacobian-Vector Product. To enhance robustness, we adopt an adaptive mean flow loss $\mathcal{L}_{flow} = w \cdot \|u_\theta - \mathrm{sg}(u_t)\|^2$, where the weight $w$ suppresses outliers.

To ground the latent trajectory rectification within an interpretable physical domain, we further introduce semantic-aware Manifold-GVC, where GVC denotes Geometric Velocity Consistency. Given the Mean Flow error $\delta u = u_\theta - \mathrm{sg}(u_t)$, this module maps its local effect to the FLAME vertex space and constrains region-specific geometric deviations. We translate latent updates into observable mesh deformations $\delta V$ via a Jacobian-aware gradient approximation:

$$\delta V \approx \frac{\mathcal{V}(x_{base} + \epsilon \cdot \delta u) - \mathcal{V}(x_{base})}{\epsilon} \tag{9}$$

where $\mathcal{V}$ denotes the FLAME decoder, $x_{\text{base}} = \mathrm{sg}(x_1)$ is the detached clean-motion anchor used to locally linearize the decoder, and $\epsilon$ is a small perturbation scale. Since speech-driven motion exhibits region-dependent dynamics where lips dominate articulation while other-face regions encode style, we implement a category-wise semantic alignment strategy:

$$\mathcal{L}_{gvc} = \lambda_{lip}\mathcal{L}^{lip}_{gvc} + \lambda_{str}\mathcal{L}^{str}_{gvc} \tag{10}$$

Specifically, we allocate stronger constraints to articulation-critical regions ($\mathcal{L}^{lip}_{gvc}$) and moderate constraints to style-relevant structural context ($\mathcal{L}^{str}_{gvc}$). This manifold rectification helps preserve articulation-critical signals from being weakened by stochastic variations in the generation trajectory, enabling more precise and expressive one-step inference.

### 3.4 Optimization

We adopt a progressive training strategy to transition the model from modeling instantaneous velocities to achieving global trajectory straightening.

**Stage I: Instantaneous Velocity Learning.** We first initialize the network to approximate the conditional vector field using standard Flow Matching. In this stage, the reference time is fixed as $r = t$, forcing the DTDT to learn the instantaneous velocity $v_\theta$ of the optimal transport path. The objective for this initial phase is:

$$\mathcal{L}_{StageI} = \mathcal{L}_{cfm} + \mathcal{L}_{geo} \tag{11}$$

where $\mathcal{L}_{cfm} = \|v_\theta - (x_1 - x_0)\|^2$. This stage ensures the model captures fundamental cross-modal correlations and establishes a high-quality motion manifold through basic geometric supervision.

**Stage II: Mean Flow Rectification.** Building upon the initial weights, we proceed to the trajectory rectification phase. To ensure a stable transition, the final layer of the $\text{MLP}_r$ is zero-initialized. We enable the interleaved switch with a curriculum sampling strategy: in each iteration, we assign $r = t$ with 50% probability to consolidate instantaneous dynamics, and sample $r \sim \mathcal{U}[0, 1]$ for the remaining 50% to learn the average velocity field $u_\theta$. The total objective is formulated as:

$$\mathcal{L}_{StageII} = \mathcal{L}_{flow} + \mathcal{L}_{gvc} + \mathcal{L}_{geo} \tag{12}$$

where $\mathcal{L}_{flow}$ is the adaptive Mean Flow loss supervised by the JVP-rectified target $u_t$. Crucially, we incorporate Semantic-aware Manifold-GVC ($\mathcal{L}_{gvc}$) to project latent trajectory updates onto the interpretable surface manifold for region-specific shaping. This joint optimization of geometry and trajectory enables MeanTalker to achieve robust one-step inference with precise and synchronized facial motion.

## 4 Experiments

### 4.1 Setup

**Datasets.** We evaluate our proposed MeanTalker on the public TFHP dataset [26], a comprehensive benchmark specifically designed for 3D facial animation. This dataset comprises 1,052 high-definition video clips collected from diverse scenarios, totaling approximately 26.5 hours of footage from 588 distinct subjects. All videos are processed at 25 fps to extract accurate FLAME parameters, providing reliable ground truth for geometry reference. We strictly follow the official protocol to partition the dataset, utilizing 460 subjects for training and 64 subjects for validation. The testing set contains 64 unseen subjects with 120 video sequences, amounting to a total duration of 3.15 hours. This extensive test set features a rich diversity of vocal timbres, ages, and genders, offering a rigorous environment to assess the model's generalization capabilities across different identities.

**Implementation Details.** We implement MeanTalker using PyTorch and train it on a single NVIDIA A100 GPU (80GB). The spatial dimension of the motion latent space corresponds to the 5023 vertices of the FLAME topology. The model is optimized using the Adam optimizer with a learning rate of $1 \times 10^{-4}$. We employ a gradual warm-up scheduler for the first 5,000 iterations to stabilize the training dynamics. For temporal context, each training sample is formed from two consecutive 100-frame clips, where a 10-frame historical motion window from the previous clip is used to smooth the transition to the target clip.

**Baselines.** We compare MeanTalker with five representative state-of-the-art methods: **MultiTalk** [27], **ScanTalk** [19], **DiffPoseTalk** [26], **ARTalk** [3], and **MSMD** [20]. MultiTalk utilizes a transformer-based architecture with discrete motion priors, while ScanTalk emphasizes geometric consistency across varying mesh topologies. DiffPoseTalk and MSMD adopt diffusion models for stylistic

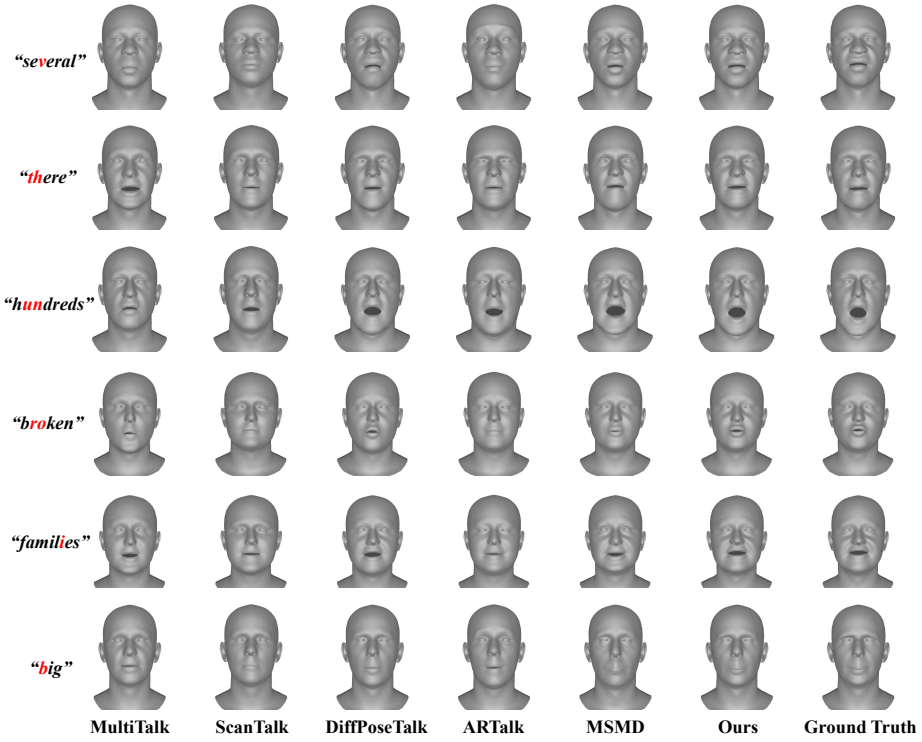


**Fig. 3: Qualitative comparison of synthesized 3D facial animations.** We compare MeanTalker with state-of-the-art methods across various challenging phonetic articulations. Our method synthesizes highly accurate lip movements (e.g., the pronounced mouth opening in "hundreds" and "big") and preserves expressive facial details (e.g., "families"), closely matching the ground truth without exhibiting the over-smoothed or misaligned artifacts observed in baselines.

facial motion generation, with MSMD further introducing style-basis conditioning. ARTalk employs multi-scale vector quantization to model temporal motion dynamics. For evaluation, all baselines are tested on the same TFHP test split with identical metrics. We use the closest publicly available setting for each method, including released checkpoints when available and training with official code when needed. More detailed settings are provided in the supplementary material.

### 4.2 Quantitative Evaluation

**Evaluation Metrics.** Following the evaluation protocols widely adopted in previous state-of-the-art studies [3,5,26], we employ a comprehensive set of metrics to quantitatively assess synchronization accuracy, motion naturalness, and stylistic expressiveness. To quantify the precision of lip-speech synchronization, we report the Lip Vertex Error (LVE) [22], which calculates the maximum $L_2$

**Table 1:** Quantitative comparison with state-of-the-art methods on the TFHP benchmark test set. The best results are **bolded** and the second best are underlined. Inference latency (RTF) is evaluated on a single NVIDIA RTX 3090 GPU.

| **Methods** | **Paradigm** | **LVE** ↓ (mm) | **FDD** ↓ ($\times 10^{-5}$ m) | **MOD** ↓ (mm) | **MVE** ↓ (mm) | **RTF** ↓ |
|---|---|---|---|---|---|---|
| MultiTalk [27] | Autoregressive | 11.099 | 8.348 | 3.853 | 1.494 | 0.288 |
| ScanTalk [19] | Regression | 13.145 | 12.550 | 4.365 | 2.254 | 0.049 |
| DiffPoseTalk [26] | Diffusion | 9.119 | 5.458 | 3.225 | 0.938 | 0.919 |
| ARTalk [3] | Autoregressive | 11.783 | 12.049 | 4.057 | 1.643 | 0.020 |
| MSMD [20] | Diffusion | 8.390 | 5.763 | 2.982 | 0.829 | 1.015 |
| **MeanTalker** (Ours) | Mean Flow | **7.879** | **5.189** | **2.830** | **0.788** | **0.004** |

error of lip region vertices for each frame. For a holistic assessment of facial geometry reconstruction, we calculate the Mean Vertex Error (MVE) by averaging the coordinate distance across all 5,023 facial vertices. Moving beyond basic lip sync, we utilize the Upper-face Dynamics Deviation (FDD) [32] to evaluate the naturalness and richness of facial expressions. This metric computes the difference in motion standard deviation for upper-face vertices between the generated and ground-truth sequences. Regarding stylistic articulation, we adopt the Mouth Opening Difference (MOD) proposed in DiffPoseTalk [26]. Given that the specific vertex definitions for this metric were not publicly detailed, we standardize the calculation by computing the average vertical distance between three validated pairs of inner lip seam vertices (FLAME indices: Top $\{3533, 2785, 1668\}$ and Bottom $\{3513, 2929, 1827\}$), which effectively captures the intensity of mouth opening variations. Finally, to evaluate computational efficiency, we report the Real-Time Factor (RTF), defined as the average inference time required to generate facial motions for one second of input audio on a single RTX 3090 GPU.

**Results.** Tab. 1 summarizes the quantitative comparisons on the TFHP benchmark. MeanTalker consistently outperforms existing baselines across all evaluated metrics. For lip-sync precision, our method achieves an exceptional LVE of 7.88 mm, significantly outperforming the recent diffusion baseline MSMD (8.39 mm). Regarding global facial kinematics, MeanTalker yields the lowest MVE (0.79 mm) and FDD ($5.19 \times 10^{-5}$ m), guaranteeing accurate and stable dynamic reconstruction. Notably, MeanTalker achieves these top-tier metrics with an exceptional Real-Time Factor (RTF) of 0.004, accelerating generation by over 200× compared to iterative diffusion models (e.g., DiffPoseTalk). When including HuBERT feature extraction and FLAME-to-vertex conversion, the full pipeline still costs only about 0.0048s per second of audio. This effectively bridges the gap between precise motion generation and real-time deployment.

**Table 2:** Quantitative ablation study. **Pre-T** represents the pretrained Standard FM model (NFE=25). **w/o Rectification** is forced to run in one step (NFE=1).

| Variant | Paradigm | LVE ↓ | FDD ↓ | MOD ↓ | MVE ↓ |
|---|---|---|---|---|---|
| Pre-T (Standard FM) | Coupled ($t \equiv r$) | 8.906 | **4.778** | 3.155 | 0.918 |
| w/o Rectification | Coupled ($t \equiv r$) | 9.189 | 6.511 | 3.231 | 0.940 |
| w/o Style Code | Full Stage | 9.144 | 5.919 | 3.076 | 0.930 |
| w/o Progressive Init | Direct Stage-II | 8.272 | 5.977 | 3.133 | 0.810 |
| w/o GATL | MSE Only | 8.735 | 5.152 | 3.187 | 0.874 |
| **MeanTalker (Full)** | Full Stage | **7.879** | 5.189 | **2.830** | **0.788** |

### 4.3 Qualitative Evaluation

Fig. 3 provides a qualitative comparison of generated facial meshes across various challenging phonetic articulations. Visual results demonstrate that MeanTalker yields the most accurate lip formations and highest geometric expressiveness, closely mirroring the GT (Ground Truth). Specifically, baselines like MultiTalk and ScanTalk frequently produce under-articulated lip movements, failing to capture the required mouth amplitude for wide vowels (e.g., the "u" in "hundreds") or the tight closure for bilabial consonants (e.g., the "b" in "big"). Furthermore, while generative baselines such as ARTalk, DiffPoseTalk, and MSMD show improved lip shapes, they tend to generate overly smoothed facial surfaces that lack dynamic muscular tension. For instance, when pronouncing the "i" in "families", these baselines struggle to fully capture the pronounced nasolabial folds (smile lines) and cheek deformations present in the GT. In contrast, MeanTalker strictly aligns with the GT in both precise lip rounding (e.g., the "o" in "broken") and rich facial details, faithfully reconstructing the local geometric variations induced by active speech.

### 4.4 Ablation Study

To rigorously validate the effectiveness of our proposed framework, we evaluate MeanTalker against several degraded variants, with quantitative and qualitative results detailed in Tab. 2 and Fig. 4, respectively. First, to assess the impact of the Dual-Time Denoising Transformer (DTDT), we evaluate the pretrained Standard FM model directly using a single step (NFE=1), denoting this variant as **w/o Rectification**. In this setting, the simulation and trajectory times remain coupled ($r \equiv t$). As visually demonstrated in Fig. 4, forcing standard flow matching into one-step inference causes severe global geometric distortions extending across the cheeks and brow ridge, verifying that its curved latent trajectory cannot be traversed instantly without our dual-time encoding mechanism. Next, we investigate the contribution of our dual-space physical rectification (**w/o GATL**) by performing trajectory learning using only the latent flow matching loss (MSE), thereby discarding all explicit geometric constraints. Depicted in the

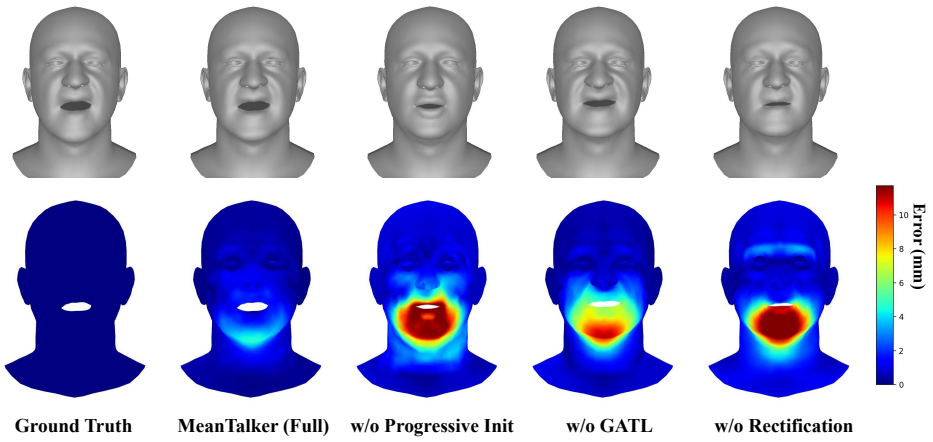


**Fig. 4: Qualitative ablation of MeanTalker.** *Top row*: Meshes highlighting geometric surface details. *Bottom row*: Vertex error heatmaps relative to the Ground Truth. Omitting the progressive curriculum (**w/o Progressive Init**) or Geometric-Aware Trajectory Learning (**w/o GATL**) degrades phonetic articulation, resulting in severe localized lip-sync errors. Forcing standard flow matching into one-step inference (**w/o Rectification**) causes global geometric distortions extending across the cheeks. In contrast, our full MeanTalker precisely captures the target facial motion with minimal spatial error.

error heatmaps (Fig. 4), this omission leads to under-articulated mouth shapes and severe localized lip-sync errors, confirming that GATL is indispensable for preserving high-frequency facial details and precise lip closures during trajectory straightening. Furthermore, training the framework from scratch without the pre-trained standard FM weights (**w/o Progressive Init**) not only leads to a clear degradation in quantitative spatial metrics and visible articulation errors, but also delays convergence (approximately 65K vs. 30K iterations). Finally, we also ablate the temporal style encoder (w/o Style Code), where the subsequent drop in the FDD metric confirms its essential role in capturing the global dynamics of the speaker.

### 4.5 User Study

To perceptually evaluate the visual quality and realism of the generated 3D facial motions, we conducted a comprehensive user study comparing MeanTalker against other strong baselines: MultiTalk [27], ScanTalk [19], DiffPoseTalk [26], ARTalk [3], and MSMD [20]. Participants were shown the generated animations alongside the corresponding ground-truth videos as references and asked to assign a Mean Opinion Score (MOS) on a scale of 1 to 5 across three dimensions: lip synchronization accuracy (Lip Sync), similarity to the reference speaking style (Style Sim), and overall naturalness of the facial dynamics (Natural). As reported in Tab. 3, MeanTalker achieves the highest scores on all three dimen-

**Table 3:** Quantitative results of the user study based on Mean Opinion Scores (MOS). The best results are **bolded**.

| **Methods** | **Lip Sync ↑** | **Style Sim ↑** | **Natural ↑** |
|---|---|---|---|
| MultiTalk [27] | 2.38 | 2.46 | 2.41 |
| ScanTalk [19] | 1.92 | 2.05 | 1.81 |
| DiffPoseTalk [26] | 3.60 | 3.60 | 3.69 |
| ARTalk [3] | 3.06 | 2.15 | 2.96 |
| MSMD [20] | 3.69 | 3.63 | 3.86 |
| **MeanTalker** (Ours) | **4.32** | **4.05** | **4.24** |

sions, securing 4.32 for lip synchronization, 4.05 for style similarity, and 4.24 for naturalness. With MSMD serving as the second-best performer, these perceptual ratings strongly align with our objective evaluations, further corroborating that MeanTalker produces highly precise lip movements while maintaining expressive and natural facial dynamics.

## 5 Conclusion

In this paper, we presented MeanTalker, a novel one-step speech-to-motion generative framework based on Mean Flow. By synergizing a Dual-Time Denoising Transformer (DTDT) for stable progressive velocity modeling and Geometric-Aware Trajectory Learning (GATL) for dual-space physical rectification, our method achieves expressive and natural facial animations without the latency of iterative sampling. Extensive experiments demonstrate that MeanTalker delivers superior lip synchronization accuracy compared to existing baselines, while achieving an exceptional Real-Time Factor (RTF) of $\approx 0.004$. We believe this highly efficient flow-based paradigm provides a robust foundation for real-time speech-driven animation applications, and future work will explore its extension to explicit emotion control, full-body avatar animation, and gesture generation.

## Acknowledgements

This work was supported by the Major Key Project of Pengcheng Laboratory (Grant No. PCL2025A17-1). The authors gratefully acknowledge this support.

## References

1. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 157–164 (2023)
2. Chen, P., Wei, X., Lu, M., Chen, H., Tian, F.: Diffusiontalker: Efficient and compact speech-driven 3d talking head via personalizer-guided distillation. arXiv preprint arXiv:2503.18159 (2025)

3. Chu, X., Goswami, N., Cui, Z., Wang, H., Harada, T.: Artalk: Speech-driven 3d head animation via autoregressive model. arXiv preprint arXiv:2502.20323 (2025)
4. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, learning, and synthesis of 3d speaking styles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10101–10111 (2019)
5. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: Faceformer: Speech-driven 3d facial animation with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18770–18780 (2022)
6. Frans, K., Hafner, D., Levine, S., Abbeel, P.: One step diffusion via shortcut models. In: International Conference on Learning Representations. vol. 2025, pp. 34668–34684 (2025)
7. Geng, Z., Deng, M., Bai, X., Kolter, Z., He, K.: Mean flows for one-step generative modeling. Advances in Neural Information Processing Systems **38**, 75460–75482 (2026)
8. Han, Z., Teye, M., Yadgaroff, D., Bütepage, J.: Tiny is not small enough: High quality, low-resource facial animation models through hybrid knowledge distillation. ACM Transactions on Graphics (TOG) **44**(4), 1–18 (2025)
9. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al.: Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014)
10. Haque, K.I., Yumak, Z.: Facexhubert: Text-less speech-driven e (x) pressive 3d facial animation synthesis using self-supervised speech representation learning. In: Proceedings of the 25th International Conference on Multimodal Interaction. pp. 282–291 (2023)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
12. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
13. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6), 194–1 (2017)
14. Li, X., Liu, J., Liang, Y., Niu, Z., Chen, W., Chen, X.: Meanaudio: Fast and faithful text-to-audio generation with mean flows. arXiv preprint arXiv:2508.06098 (2025)
15. Lin, Y., Fan, Z., Wu, X., Xiong, L., Peng, L., Li, X., Kang, W., Lei, S., Xu, H.: Glditalker: Speech-driven 3d facial animation with graph latent diffusion transformer. arXiv preprint arXiv:2408.01826 (2024)
16. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
17. 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)
18. Morishima, S.: Real-time talking head driven by voice and its application to communication and entertainment. In: AVSP. vol. 98, pp. 195–199 (1998)
19. Nocentini, F., Besnier, T., Ferrari, C., Arguillere, S., Berretti, S., Daoudi, M.: Scantalk: 3d talking heads from unregistered scans. In: European Conference on Computer Vision. pp. 19–36. Springer (2024)
20. Pan, Y., Singh, K., Hafemann, L.G.: Model see model do: Speech-driven facial animation with style control. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Papers. pp. 1–10 (2025)
21. Peng, Z., Wu, H., Song, Z., Xu, H., Zhu, X., He, J., Liu, H., Fan, Z.: Emotalk: Speech-driven emotional disentanglement for 3d face animation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 20687–20697 (2023)

22. Richard, A., Zollhöfer, M., Wen, Y., De la Torre, F., Sheikh, Y.: Meshtalk: 3d face animation from speech using cross-modality disentanglement. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1173–1182 (2021)
23. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023)
24. Stan, S., Haque, K.I., Yumak, Z.: Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In: Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games. pp. 1–11 (2023)
25. Sun, X., Zhang, L., Zhu, H., Zhang, P., Zhang, B., Ji, X., Zhou, K., Gao, D., Bo, L., Cao, X.: Vividtalk: One-shot audio-driven talking head generation based on 3d hybrid prior. In: 2025 International Conference on 3D Vision (3DV). pp. 713–722. IEEE (2025)
26. Sun, Z., Lv, T., Ye, S., Lin, M., Sheng, J., Wen, Y.H., Yu, M., Liu, Y.j.: Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. ACM Transactions on Graphics (TOG) **43**(4), 1–9 (2024)
27. Sung-Bin, K., Chae-Yeon, L., Son, G., Hyun-Bin, O., Ju, J., Nam, S., Oh, T.H.: Multitalk: Enhancing 3d talking head generation across languages with multilingual video dataset. arXiv preprint arXiv:2406.14272 (2024)
28. Taylor, S.L., Mahler, M., Theobald, B.J., Matthews, I.: Dynamic units of visual speech. In: Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation. pp. 275–284 (2012)
29. 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. Springer (2020)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
31. Wu, S., Haque, K.I., Yumak, Z.: Probtalk3d: Non-deterministic emotion controllable speech-driven 3d facial animation synthesis using vq-vae. In: Proceedings of the 17th ACM SIGGRAPH conference on motion, interaction, and games. pp. 1–12 (2024)
32. Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., Wong, T.T.: Codetalker: Speech-driven 3d facial animation with discrete motion prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12780–12790 (2023)
33. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8652–8661 (2023)
34. Zhou, L., Ermon, S., Song, J.: Inductive moment matching. arXiv preprint arXiv:2503.07565 (2025)
35. Zhou, Y., Xu, Z., Landreth, C., Kalogerakis, E., Maji, S., Singh, K.: Visemenet: Audio-driven animator-centric speech animation. ACM Transactions on Graphics (ToG) **37**(4), 1–10 (2018)