PointTalk: Audio-Driven Dynamic Lip Point Cloud for 3D Gaussian-based Talking Head Synthesis

Yifan Xie[1,2], Tao Feng[1], Xin Zhang[1], Xiangyang Luo[1], Zixuan Guo[3], Weijiang Yu[4], Heng Chang[5], Fei Ma[1*], Fei Richard Yu[6,7]

[1]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[2]Xi'an Jiaotong University
[3]Peking University
[4]Sun Yat-sen University
[5]Tsinghua University
[6]Shenzhen University
[7]Carleton University
{ivanxie416, fengtao0127, zhangx0526, goodluoxy, gzx2019, weijiangyu8}@gmail.com, changh17@tsinghua.org.cn, mafei@gml.ac.cn, richard.yu@ieee.org

Abstract

Talking head synthesis with arbitrary speech audio is a crucial challenge in the field of digital humans. Recently, methods based on radiance fields have received increasing attention due to their ability to synthesize high-fidelity and identity-consistent talking heads from just a few minutes of training video. However, due to the limited scale of the training data, these methods often exhibit poor performance in audio-lip synchronization and visual quality. In this paper, we propose a novel 3D Gaussian-based method called PointTalk, which constructs a static 3D Gaussian field of the head and deforms it in sync with the audio. It also incorporates an audio-driven dynamic lip point cloud as a critical component of the conditional information, thereby facilitating the effective synthesis of talking heads. Specifically, the initial step involves generating the corresponding lip point cloud from the audio signal and capturing its topological structure. The design of the dynamic difference encoder aims to capture the subtle nuances inherent in dynamic lip movements more effectively. Furthermore, we integrate the audio-point enhancement module, which not only ensures the synchronization of the audio signal with the corresponding lip point cloud within the feature space, but also facilitates a deeper understanding of the interrelations among cross-modal conditional features. Extensive experiments demonstrate that our method achieves superior high-fidelity and audio-lip synchronization in talking head synthesis compared to previous methods.

Introduction

With the development of the audio-visual industry, audio-driven talking head synthesis has become crucial in computer vision and multimedia. The goal is to create videos where a person's face moves naturally in sync with input audio while maintaining their visual identity. This task has matured into a prominent research topic recently and holds potential for integration into diverse applications, including virtual avatars (Thies et al. 2020), film making (Zhang et al. 2021), and online meetings (Kim et al. 2018). Researchers have investigated various methods to tackle this task. While various 2D-based studies (Prajwal et al. 2020; Wang et al. 2023b; Zhang et al. 2023b,a) successfully synthesize talking head video using generative models, the absence of a unified facial model results in shortcomings in identity preservation and pose control. Other methods (Ji et al. 2021; Xing et al. 2023) model an individual by utilizing explicit facial structural priors, including landmarks and meshes. However, accumulated errors in these intermediate representations can significantly impact the final outcomes.

In existing research, Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) have become popular for their ability to render realistic and 3D-consistent images from novel viewpoints, playing a crucial role in synthesizing talking head videos. AD-NeRF (Guo et al. 2021) pioneered the use of conventional audio processing techniques to generate conditional features for neural radiance fields, though it faced challenges with slow training and inference speeds. In contrast, RAD-NeRF (Tang et al. 2022) and ER-NeRF (Li et al. 2023) introduced architectural innovations to streamline the enhancement of conditional features, achieving real-time inference speeds. Simultaneously, GeneFace (Ye et al. 2022) and its variant (Ye et al. 2023) have employed facial landmarks as conditions to improve the generalization capabilities for out-of-domain audio. While these methods maintain identity consistency, limited training data leads to issues with lip sync, facial details, and overall stability, reducing the realism of generated talking heads.

3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) has recently achieved impressive results in 3D scene reconstruction. It utilizes 3D Gaussians as discrete geometric primitives, resulting in a clear representation of the scene and optimized real-time rendering performance. Compared to NeRF, 3DGS not only significantly improves rendering efficiency and visual quality, but also its paradigm based on 3D Gaussians is also easier to control. This potential ease of control makes it feasible to intuitively manipulate facial movements.

*Corresponding author.

One intuitive method is to drive a Gaussian point cloud for facial motion using parametric 3D facial models (Wang et al. 2023a; Qian et al. 2024). By binding the Gaussians to the model's geometric topology, dynamic talking heads can be generated by synchronizing the displacement of the Gaussians with changes in the audio parameters.

In this paper, we propose PointTalk, a novel 3D Gaussian-based method, that attempts to utilize the 3DGS to achieve realistic and effective talking head synthesis. Unlike previous methods that directly utilize audio signals or landmarks as conditions, PointTalk uses audio to generate dynamic lip point clouds, which work together with the audio signal to create more effective talking head. Specifically, the first step involves generating a dynamic lip point cloud based on the audio signal, which captures its topological structure through a multi-resolution hash grid. Considering the dynamic nature of the whole process, simply capturing the features of each point cloud frame is insufficient for effectively guiding dynamic scenes. As a result, we develop a dynamic difference encoder to more accurately capture the nuances of lip movement. Furthermore, the compression of high-dimensional conditions into significantly lower dimensions, as observed in HyperNeRF (Park et al. 2021), results in a substantial loss of informational content. Therefore, we propose an audio-point enhancement module that synchronizes audio signals with point clouds and understands cross-modal feature correlations. The enhanced features produce optimal rendering results, generating high-quality syntheses. Experiments show that PointTalk renders visually realistic talking heads with accurate audio-lip synchronization.

In summary, the main contributions of our work are as follows:

- We propose a novel 3D Gaussian-based framework called PointTalk, which incorporates an audio-driven dynamic lip point cloud as an additional condition to achieve realistic talking head synthesis.
- We introduce an Audio2Point module for generating a dynamic lip point cloud from speech audio. Additionally, we utilize a dynamic difference encoder to more accurately capture the nuances of lip movement.
- We design an audio-point enhancement module that synchronizes audio signals with their corresponding lip point clouds and comprehends the correlation between cross-modal conditional features.

Related Work

Talking Head Synthesis

Talking head synthesis aims to create a video of a speaking person that accurately represents their identity and is perfectly synchronized with the driven audio. It consists of two main directions: 2D-based and 3D-based methods.

2D-Based Talking Head Synthesis. Some methods (Prajwal et al. 2020; Wang et al. 2023b; Zhong et al. 2023; Zhang et al. 2023b) rely on images that focus primarily on the face, particularly the mouth, to ensure the audio matches the lip movements. For example, Wav2Lip (Prajwal et al. 2020) incorporates a lip synchronization expert to oversee the accuracy of lip movements, while TalkLip (Wang et al. 2023b) utilizes a lip-reading expert to enhance the clarity and precision of these movements. However, since these methods reconstruct the lips using only a few reference frames, they struggle to maintain consistent identity. Recently, diffusion models (Ho, Jain, and Abbeel 2020) have been employed to enhance lip-sync and image quality (Shen et al. 2023b; Ma et al. 2023), but they tend to be slow during inference. Additionally, there are some other methods (Ji et al. 2022; Zhang et al. 2023a; Ye et al. 2024) that require only a single image to generate a dynamic talking head video. For instance, EAMM (Ji et al. 2022) generates realistic emotional talking faces from a single shot. Real3D-Portrait (Ye et al. 2024) enhances one-shot 3D avatar reconstruction and talking face animation. However, such methods make it difficult to generate natural head poses and facial expressions, resulting in unrealistic visual representations.

3D-Based Talking Head Synthesis. Conventional 3D-based methods (Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017; Ji et al. 2021) often utilize 3D Morphable Models (3DMM) (Blanz and Vetter 2023) for talking head synthesis. However, the use of intermediate representations can lead to the accumulation of errors. With the recent rise of Neural Radiance Fields (NeRF) (Mildenhall et al. 2021), it has been applied to tackle 3D head structure problems in audio-driven talking head synthesis (Guo et al. 2021; Tang et al. 2022; Li et al. 2023; Ye et al. 2023; Shen et al. 2023a; Peng et al. 2024). AD-NeRF (Guo et al. 2021) stands as the pioneering method in utilizing NeRF for audio-driven talking head synthesis. By incorporating Instant-NGP (Müller et al. 2022), RAD-NeRF (Tang et al. 2022) has achieved significant enhancements in both visual quality and efficiency. ER-NeRF (Li et al. 2023) introduces a triple-plane hash encoder designed to eliminate empty spatial regions and generates region-aware conditional features through an attention mechanism. GeneFace (Ye et al. 2022) and its variant (Ye et al. 2023) generate content conditioned on estimated facial landmarks. SyncTalk (Peng et al. 2024) advances the realism of audio-driven talking head videos by accurately synchronizing facial identity, lip movements, expressions, and head poses. Despite its excellent performance, the inference process is not real-time. TalkingGaussian (Li et al. 2024) first attempts to utilize the 3DGS (Kerbl et al. 2023) to address the facial distortion problem in existing radiance-fields-based methods. This paper introduces a 3D Gaussian-based method that significantly enhances visual quality and audio-lip synchronization. Moreover, it ensures that the inference process operates in real time.

Point Cloud Learning

Point cloud learning (Qi et al. 2017a; Wang et al. 2019; Xie et al. 2023, 2024) is a crucial research area in 3D computer vision. PointNet (Qi et al. 2017a) utilizes point-wise MLPs and pooling layers to aggregate features for understanding 3D scenes. PointNet++ (Qi et al. 2017b) advances PointNet (Qi et al. 2017a) by incorporating hierarchical sampling strategies. Point-BERT (Yu et al. 2022) adapts the masked language modeling approach of BERT (Kenton and Toutanova 2019) to the 3D realm. RECON (Qi et al. 2023) combines the strengths of generative and contrastive

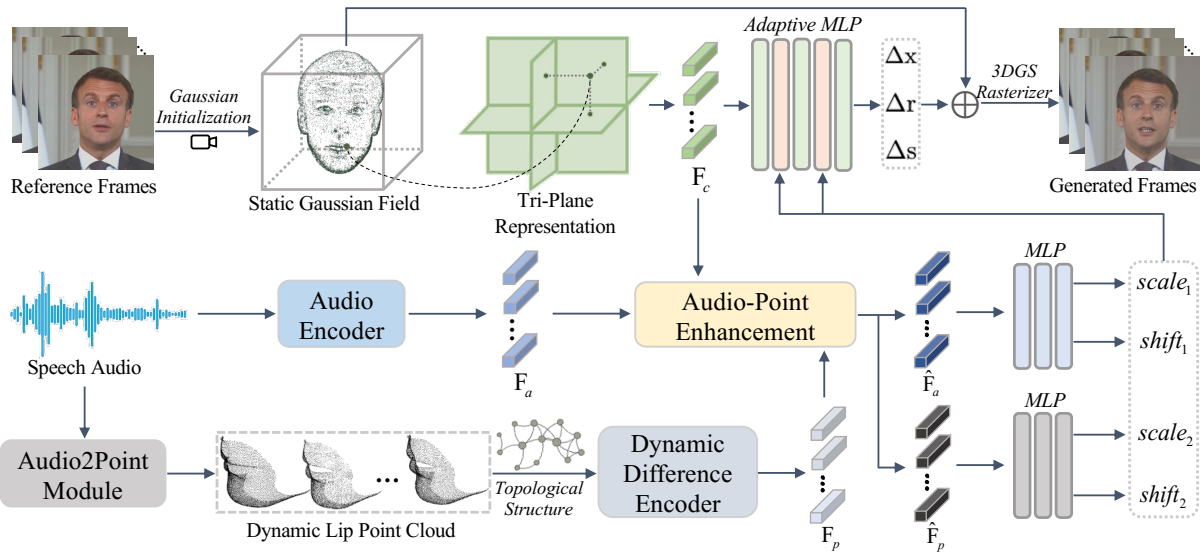

Figure 1: Overview of PointTalk. Utilizing the static Gaussian field to optimize the coarse Gaussian head from a random point cloud. Then the tri-plane encoder and audio encoder independently extract the spatial geometry feature F_c and the audio feature F_a. The Audio2Point module generates a dynamic lip point cloud based on the input audio signals. Subsequently, the dynamic lip point cloud's topological structure is established, and the dynamic difference encoder extracts the point cloud feature F_p. Moreover, the audio-point enhancement module synchronizes the audio signals with the point cloud to facilitate information interaction, thereby obtaining the enhancement features $\hat{\mathrm{F}}_a$ and $\hat{\mathrm{F}}_p$. Ultimately, the enhancement features are fed into two MLP decoders to compute the scale and shift factors. By integrating these factors with F_c, an adaptive MLP is deployed to predict the deformation parameters for 3DGS rasterizer.

learning paradigms for enhanced 3D representation learning. PointGPT (Chen et al. 2024) extends the concept of GPT to point clouds. Our method generates a dynamic lip point cloud from audio and captures its topological structure using a multi-resolution hash grid.

Method

In this section, we introduce the proposed PointTalk, as illustrated in Figure 1. PointTalk is structured around three key components: 1) The *Multi-Attribute Branches* are utilized to process 3D Gaussians, audio signals and dynamic lip point clouds separately. 2) The *Audio-Point Enhancement* module, which not only synchronizes the audio signals with the point cloud but also enhances information interaction. 3) The *Adaptive 3DGS Rendering* can render the final talking head video effectively. We will delve into the specifics of these components and the associated loss functions in the subsequent subsections.

Multi-Attribute Branches

3D Gaussian Branch. The static Gaussian field preserve the Gaussian primitives with the canonical parameters. Specifically, a Gaussian primitive can be described with a position $\mathrm{x} \in \mathbb{R}^3$, a rotation quaternion $\mathrm{r} \in \mathbb{R}^4$, a scaling factor $\mathrm{s} \in \mathbb{R}^3$, an opacity value $\alpha \in \mathbb{R}^1$, and a d-dimensional color feature $f \in \mathbb{R}^d$. Consequently, the i-th Gaussian primitive $\mathcal{G}_i$ keeps a set of parameters $\theta_i = \{\mathrm{x}_i, \mathrm{r}_i, \mathrm{s}_i, \alpha_i, f_i\}$ and the formulation can be described as:

$$\mathcal{G}_i(p) = e^{-\frac{1}{2}(p-\mathrm{x}_i)^T \Sigma_i^{-1}(p-\mathrm{x}_i)}, \tag{1}$$

where the covariance matrix Σ can be obtained using r and s. We first initialize the talking head with the static Gaussian field by the training video frames to get a coarse Gaussian head. Then, we learn the deformation parameters to deform the Gaussian head with the audio.

Although Gaussian primitives are effective in representing the Gaussian head, they lack a regional position encoding due to their explicit structure. To address this, we propose utilizing a tri-plane representation (Chan et al. 2022). Specifically, for a given position $\mathrm{x} = (x, y, z) \in \mathbb{R}^{\mathrm{XYZ}}$, it is transformed through an encoding process where its projected values are utilized by three distinct 2D multi-resolution hash encoders (Müller et al. 2022):

$$\mathcal{H}^{\mathrm{AB}} : (a, b) \rightarrow \mathrm{f}_{ab}^{\mathrm{AB}}, \tag{2}$$

where the output $\mathrm{f}_{ab}^{\mathrm{AB}} \in \mathbb{R}^{LD}$, defined by L levels and D feature dimensions per entry, represents the planar geometric feature associated with the projected coordinate (a, b). In this scenario, $\mathcal{H}^{\mathrm{AB}}$ refers to the multi-resolution hash encoder for plane $\mathbb{R}^{\mathrm{AB}}$. By concatenating these outputs, we obtain the final spatial geometry feature $\mathrm{f}_c \in \mathbb{R}^{3LD}$, which integrates the encoded geometric information. The whole process can be described as:

$$\mathrm{f}_c = \mathcal{H}^{\mathrm{XY}}(x, y) \oplus \mathcal{H}^{\mathrm{YZ}}(y, z) \oplus \mathcal{H}^{\mathrm{XZ}}(x, z), \tag{3}$$

where $\oplus$ denotes the concatenation operator.

Audio Branch. In audio branch, we use an automatic speech recognition (ASR) module (Amodei et al. 2016; Hsu et al. 2021) to extract audio features $\mathrm{F}_a \in \mathbb{R}^{T \times F}$ from the audio

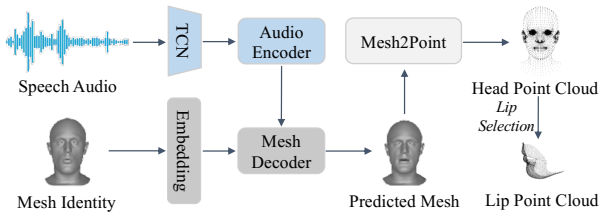

Figure 2: The pipeline of the Audio2Point module.

track, where T denotes the frame count and F represents the feature dimension. For additional details, please refer to the supplementary materials.

Lip Point Cloud Branch. Inspired by mesh-based methods (Fan et al. 2022; Peng et al. 2023), we design an Audio2Point module to capture an additional dynamic lip point cloud. The pipeline of the Audio2Point module is depicted in Figure 2. Specifically, the mesh is used as the reference embedding, while the speech audio is compressed by a temporal convolutional network (TCN) and an audio encoder. Then, the predicted mesh is generated by combining the audio features with the embedding. After that, the vertices of the mesh are extracted to generate the head point cloud, which is further refined to produce the final lip point cloud $P \in \mathbb{R}^{T\times N\times 3}$, where T represents the frame count and N signifies the quantity of points. For additional details, please refer to the supplementary materials.

For each frame of the point cloud, it is essential to capture its topological structure to achieve a robust representation. To this end, we employ E^3_{point}, a multi-resolution hash grid (Müller et al. 2022) that encodes the point cloud efficiently. We opt for a hash grid over a tri-plane representation when processing the lip point cloud for a key reason: the Gaussian primitives usually reach up to tens of thousands, in contrast to the lip point cloud points, which are typically in the mere hundreds. This disparity significantly reduces the potential for hash collisions. The advantages of this way will be illustrated in subsequent ablation studies.

Given the dynamic nature of the synthesis process, we believe that merely capturing the features of each point cloud frame falls short in effectively guiding the generation of the talking head. Consequently, we design a dynamic difference encoder to better capture the nuances of lip movement. Specifically, the differences are taken for the features of neighboring frames, and then all the differences are concatenated together to obtain the final point cloud features $\mathrm{F}_p \in \mathbb{R}^{(T-1)\times F}$:

$$\mathrm{F}_p = \mathrm{MLP}(\mathrm{cat}\left[E^3_{\text{point}}(P_{t+1}) - E^3_{\text{point}}(P_t)\right]_1^{T-1}), \quad (4)$$

where $\mathrm{cat}[\cdot]$ denotes the feature concatenation and t represents the frame order.

Audio-Point Enhancement

We introduce the audio-point enhancement module, designed with a dual purpose: to synchronize the audio signals with the corresponding point cloud, and to understand the correlation between cross-modal features. The detailed structure of this module is depicted in Figure 3.

Drawing inspiration from (Chung and Zisserman 2017; Prajwal et al. 2020), we recognize that synchronizing the audio signals with the video signals can facilitate the generation of the talking head to a notable extent. In our method, we assert that synchronizing the audio signals with the corresponding point cloud will likewise contribute to enhanced performance. Specifically, we introduce a cross-modal contrastive learning strategy to establish the audio-point correspondences.

Given the audio features $\mathrm{F}_a \in \mathbb{R}^{T\times F}$ and point cloud frame features $\mathrm{F}'_p \in \mathbb{R}^{T\times F}$, the process of individually extracting lip point cloud features for each frame is aimed at establishing a correspondence with the audio features. For the t-th frame, our objective is to enhance the similarity between F_{at} and $\mathrm{F}'_{pt} = E^3_{\text{point}}(P_t)$, as they both correspond to the same object. Simultaneously, we strive to reduce the similarity between F_{at} and the point cloud features of non-corresponding frames. Therefore, we can construct the loss function $l(t, \mathrm{F}_a, \mathrm{F}'_p)$:

$$l(t, \mathrm{F}_a, \mathrm{F}'_p) = -\log \frac{\exp(sim(\mathrm{F}_{at}, \mathrm{F}'_{pt})/\tau)}{\sum_{\substack{k=1 \\ k\neq t}}^{T} \exp(sim(\mathrm{F}_{at}, \mathrm{F}'_{pk})/\tau)}, \quad (5)$$

where τ is the temperature factor and $sim(\cdot,\cdot)$ denotes the cosine similarity function. And the cross-modal contrastive learning loss $\mathcal{L}_{CL}$ is then formulated as:

$$\mathcal{L}_{CL} = \sum_{t=1}^{T}[l(t, \mathrm{F}_a, \mathrm{F}'_p) + l(t, \mathrm{F}'_p, \mathrm{F}_a)]. \quad (6)$$

Dynamic conditions, including audio signals and lip point clouds, play a selective role in talking head synthesis. Many previous methods (Guo et al. 2021; Tang et al. 2022) have simply assumed that these dynamic conditions uniformly influence the entire synthesis process. To comprehend the correlation between cross-modal features, lightweight external attention (Guo et al. 2022; Li et al. 2023) is employed for information interactions. Specifically, self-attention is first utilized to compress audio and point cloud features following (Tang et al. 2022). A two-layer MLP is then utilized to capture the global context of spatial geometry. Following this, we apply the Hadamard product to multiply the global context with the compressed dynamic conditions, yielding the final enhanced features. This process can be represented as follows:

$$\hat{\mathrm{F}}_a = \mathrm{MLP}_a(\mathrm{F}_c) \odot \mathrm{SA}(\mathrm{F}_a), \quad (7)$$

where $\odot$ denotes Hadamard product. A similar process is performed for the point cloud.

Additionally, given that the whole process is audio-driven and the lip point cloud is generated from the audio signals, we aim for the point cloud features to emphasize the audio-related parts more prominently. To achieve this, we process the enhanced audio features through another two-layer MLP to capture the global content of the audio. Subsequently, we utilize the Hadamard product to derive the final enhanced point cloud features.

Methods	PSNR ↑	LPIPS ↓	FID ↓	LMD ↓	LSE-D ↓	LSE-C ↑	Time	FPS
Ground Truth	N/A	0	0	0	6.897	8.275	-	-
Wav2Lip (Prajwal et al. 2020)	-	-	11.802	4.498	**7.326**	**8.363**	-	15
TalkLip (Wang et al. 2023b)	-	-	13.574	6.149	7.641	7.027	-	10
DINet (Zhang et al. 2023b)	-	-	7.597	5.712	8.572	6.358	-	17
AD-NeRF (Guo et al. 2021)	29.196	0.1241	16.274	3.680	8.842	5.877	20h	0.1
RAD-NeRF (Tang et al. 2022)	31.904	0.0722	9.651	3.197	8.179	6.043	6h	28
GeneFace++ (Ye et al. 2023)	30.417	0.0968	11.709	3.546	7.547	6.335	8h	23
ER-NeRF (Li et al. 2023)	31.959	0.0379	**6.927**	3.125	8.254	6.174	2h	30
TalkingGaussian (Li et al. 2024)	32.398	0.0355	8.385	2.967	7.825	6.516	**1h**	**90**
PointTalk	**32.770**	**0.0337**	7.331	**2.818**	7.383	7.165	**1h**	85

Table 1: **The quantitative results of the head reconstruction**. The boldface indicates the best performance and the underline represents the second-best performance. In the self-driven evaluation, since Wav2Lip, TalkLip, and DINet have access to the ground truth with the exception of the mouth region, the PSNR and LPIPS metrics are deemed inapplicable.

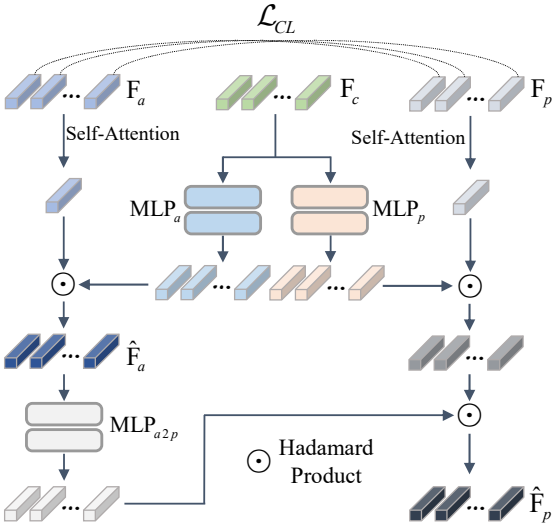

Figure 3: The detailed structure of the Audio-Point Enhancement module.

Adaptive 3DGS Rendering

Most previous methods (Guo et al. 2021; Tang et al. 2022; Li et al. 2023) typically integrate conditional features into spatial geometry features through concatenation. However, we argue that concatenation alone is insufficient for the effective fusion of cross-modal features. Inspired by Adaptive Instance Normalization (AdaIN) (Huang and Belongie 2017), we adopt a combination of a residual block and AdaIN to guide the fusion process. This process subjects the outputs to an affine transformation, incorporating both translation and scaling, parameterized by $scale_i$ and $shift_i$, respectively. These parameters are then employed to generate adaptive features F'_c:

$$\mathrm{F}'_c = (1 + scale_i) * \mathrm{F}_c + shift_i. \tag{8}$$

Ultimately, we can use the adaptive features to derive the point-wise deformation parameters $(\Delta\mathrm{x}, \Delta\mathrm{r}, \Delta\mathrm{s})$ for the 3DGS Rasterizer, which are irrelevant to the color and opacity changes. Therefore, we can calculate the final parameter $\theta_D = \{\mathrm{x} + \Delta\mathrm{x}, \mathrm{r} + \Delta\mathrm{r}, \mathrm{s} + \Delta\mathrm{s}, \alpha, f\}$ for rendering. The

	Audio A		Audio B	
Methods	LSE-D ↓	LSE-C ↑	LSE-D ↓	LSE-C ↑
Ground Truth	6.899	7.354	7.322	8.682
Wav2Lip (Prajwal et al. 2020)	**7.896**	**7.393**	**6.760**	**9.259**
TalkLip (Wang et al. 2023b)	9.013	6.229	8.705	7.528
DINet (Zhang et al. 2023b)	8.771	6.196	8.746	7.140
AD-NeRF (Guo et al. 2021)	14.432	1.274	13.896	1.877
RAD-NeRF (Tang et al. 2022)	11.639	1.941	11.082	3.135
GeneFace++ (Ye et al. 2023)	9.195	4.868	8.540	6.631
ER-NeRF (Li et al. 2023)	9.569	4.670	9.082	5.871
TalkingGaussian (Li et al. 2024)	9.171	5.327	9.061	5.745
PointTalk	8.406	6.427	8.331	7.018

Table 2: **The quantitative results of the lip synchronization**. We utilize two different audio samples to driven the same subject. The boldface indicates the best performance and the underline represents the second-best performance.

3DGS rasterizer gather N modified Gaussians with the camera information to compute the color $\mathcal{C}$ of pixel p:

$$\mathcal{C}(p) = \sum_{i \in N} c_i \hat{\alpha}_i \prod_{j=1}^{i-1} (1 - \hat{\alpha}_j), \tag{9}$$

where c_i represents the decoded color from f, $\hat{\alpha}_i$ is the result of calculating opacity α_i with the projected function. For a more comprehensive understanding of the 3DGS Rasterizer, please refer to the supplementary materials.

Loss Function

To initialize the coarse Gaussian head, we follow the original 3DGS (Kerbl et al. 2023) and utilize a combination of pixel-wise loss $\mathcal{L}_1$ and D-SSIM term $\mathcal{L}_{D-SSIM}$ (weighted by λ_1). After the initialization, we further predict the deformation parameters and input them for the 3DGS rasterizer to render the output images. Following the process of previous methods (Tang et al. 2022; Li et al. 2023), we randomly sample a set of patches from the entire image and incorporate LPIPS loss (Zhang et al. 2018) (weighted by λ_2) to improve detail resolution. This is combined with the cross-modal contrastive learning loss $\mathcal{L}_{CL}$ (weighted by λ_3) as described in Eq. 6. The overall loss function can be construsted as:

$$\mathcal{L} = \mathcal{L}_1 + \lambda_1 \mathcal{L}_{D-SSIM} + \lambda_2 \mathcal{L}_{LPIPS} + \lambda_3 \mathcal{L}_{CL}, \tag{10}$$

Figure 4: Qualitative comparison of talking head synthesis by different methods. PointTalk has the best visual effect on lip movements and facial details. Please zoom in for better visualization.

Experiments

Experimental Settings

Dataset. To ensure a fair comparison, the dataset for our experiments is sourced from (Tang et al. 2022; Ye et al. 2022; Li et al. 2023) and includes both English and French languages. We collect high-definition speaking video clips, each with an average length of approximately 7,500 frames at 25 FPS. Each raw video is cropped and resized to a resolution of 512×512, focusing on a centered portrait.

Comparison Baselines. We compare our method against three 2D-based methods, such as Wav2Lip (Prajwal et al. 2020), TalkLip (Wang et al. 2023b), and DINet (Zhang et al. 2023b), as well as five 3D-based methods, including AD-NeRF (Guo et al. 2021), RAD-NeRF (Tang et al. 2022), GeneFace++ (Ye et al. 2023), ER-NeRF (Li et al. 2023), and TalkingGaussian (Li et al. 2024).

Quantitative Evaluation

Metrics. We utilize Peak Signal-to-Noise Ratio (PSNR) to assess the overall image quality and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018) to evaluate the details. Additionally, we employ Fréchet Inception Distance (FID) (Heusel et al. 2017) to gauge image quality at the feature level. For evaluating lip synchronization, we recommend using the Landmark Distance (LMD), which quantifies the distance between lip landmarks. Furthermore, we introduce Lip Sync Error Distance (LSE-D) and Lip Sync Error Confidence (LSE-C), consistent with Wav2Lip (Prajwal et al. 2020), to assess the synchronization between lip movements and audio.

Comparison Settings. In our quantitative evaluation, we assess our method in two distinct settings: the head reconstruction setting and the lip synchronization setting. For the head reconstruction setting, we divide each video into training and test datasets to evaluate the quality of the talking head reconstruction. For the lip synchronization setting, we extract two out-of-distribution audio clips, named Audio A and Audio B. These audio clips are utilized to drive the same subject for comparison in lip synchronization.

Evaluation Results. The evaluation results of the head reconstruction setting are illustrated in Table 1. It can be observed that our image quality is superior to other methods in almost all aspects. In terms of lip synchronization, our results surpass most methods. Specifically, while one-shot methods such as Wav2Lip (Prajwal et al. 2020), Talklip (Wang et al. 2023b), and DINet (Zhang et al. 2023b) perform excellently on the LSE-D and LSE-C metrics and can synthesize talking heads without per-identity training, they score poorly on other metrics. Compared to other 3D-based methods, our method outperforms them in most metrics. Especially in lip synchronization metrics, the improvement of our method is more obvious. This is mainly due to the assistance of the dynamic lip point cloud. In terms of lip synchronization, the evaluation results are shown in Table 2. Our method also demonstrates an excellent generalization ability to synthesize lip-sync talking heads. Additionally, our method maintains superior performance in both training time and inference FPS, approximating TalkingGaussian (Li et al. 2024), which demonstrates the high efficiency of PointTalk.

Qualitative Evaluation

Evaluation Results. To more intuitively evaluate image quality and lip synchronization, we present a comparison of our method with others in Figure 4. We showcase key frames from a clip and close-up details of two talking heads. The results demonstrate that our PointTalk captures finer details

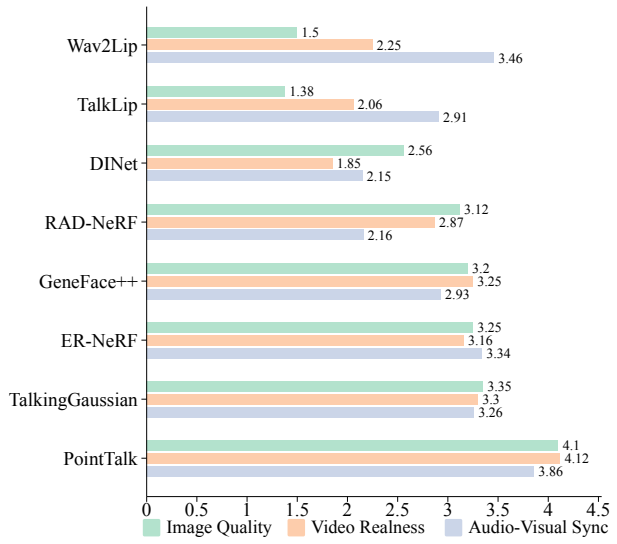

Figure 5: User study. The rating scale ranges from 1 to 5, with higher numbers indicating better performance.

and achieves the highest accuracy in lip synchronization.

User Study. To conduct a more comprehensive evaluation of PointTalk, we implement a user study questionnaire. We select 36 video clips generated during the quantitative evaluation and invite 16 participants to take part in the survey. Participants are required to rate the generated videos based on three aspects: (1) Image Quality, (2) Video Realness, and (3) Audio-Visual Synchronization. The average scores for each criterion are presented in Figure 5. PointTalk outperforms previous methods in all aspects. Notably, in terms of image quality and video realness, PointTalk exceeds the second-ranked methods by a margin of over 20%.

Ablation Study

We conduct an ablation study under the head reconstruction setting to assess the contributions of various components in our PointTalk. The results are presented in Table 3.

Lip Point Cloud Encoder. In our method, we utilize a multi-resolution hash grid to capture the topological structure of the lip point cloud. This setting is replaced with a graph neural network (GNN) (Wang et al. 2019) and tri-plane representation (Li et al. 2023), as illustrated in Table 3 (lines 1-2). The findings reveal that solely employing a GNN falls short in efficiently extracting the topological structure. Moreover, the tri-plane representation tends to induce a certain level of information loss. Further investigation into the optimal level (L) and dimension (F) of the multi-resolution hash grid is presented in Table 3 (lines 4-7). Optimal performance is achieved at $\mathrm{L} = 8$ and $\mathrm{F} = 4$, which we adopt for all our experiments. Additionally, we develop a dynamic difference encoder (DDE) to better capture the subtle movements of the lips. A comparison of the results in Table 3 (lines 8 and 11) clearly shows that the DDE enhances lip-sync metrics. Figure 6 further demonstrates the inaccuracies in lip shape that occur without the use of DDE.

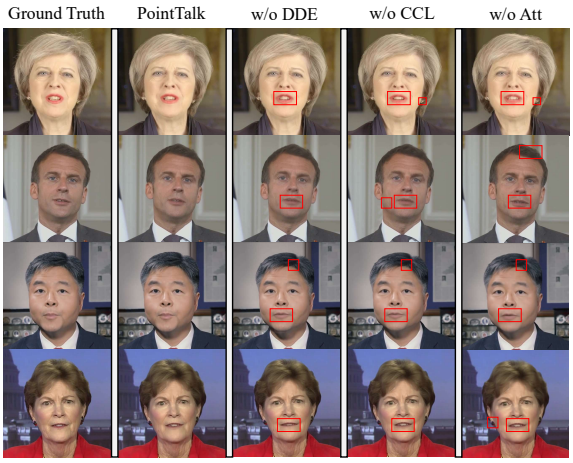

Figure 6: Visualization of the ablation study.

Setting	PSNR↑	LPIPS↓	FID↓	LMD↓	LSE-D↓	LSE-C↑
GNN	31.011	0.0416	9.353	3.655	8.135	6.253
Tri-Plane	32.390	0.0389	7.319	2.873	7.636	6.816
Hash Grid	**32.704**	**0.0365**	**7.194**	**2.773**	**7.367**	**7.079**
L=32 F=1	32.515	0.0378	7.324	2.915	7.564	6.818
L=16 F=2	32.662	0.0367	7.295	2.864	7.471	6.995
L=8 F=4	**32.704**	**0.0365**	**7.194**	**2.773**	**7.367**	**7.079**
L=4 F=8	32.617	0.0376	7.523	2.818	7.679	6.790
w/o DDE	32.032	0.0384	**6.997**	2.976	7.649	6.358
w/o CCL	32.125	0.0389	7.357	2.953	7.786	6.455
w/o Att	31.790	0.0418	8.353	3.115	8.118	6.124
PointTalk	**32.704**	**0.0365**	7.194	**2.773**	**7.367**	**7.079**

Table 3: Ablation study on different settings. Best performance is highlighted in bold.

Audio-Point Enhancement. To evaluate the Audio-Point Enhancement module's effectiveness, experiments in Table 3 (lines 9-11) show that removing cross-modal contrastive learning (CCL) causes misalignment between audio and lip points, while excluding external attention (Att) weakens cross-modal feature correlation. As shown in Figure 6, these ablations result in inaccurate lip shapes and blurred details.

Conclusion

In this paper, we propose PointTalk, a novel 3D-Gaussian based method that incorporates an audio-driven dynamic lip point cloud to achieve realistic talking head synthesis. Initially, we introduce an Audio2Point module for generating a dynamic lip point cloud and develop a dynamic difference encoder to precisely encode the nuances of lip movement. Furthermore, we integrate an audio-point enhancement module. This module not only synchronizes audio signals with their corresponding lip point clouds but also understands the correlation between cross-modal conditional features. Extensive experiments in various settings demonstrate the superior performance of PointTalk.

References

Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*, 173–182. PMLR.

Blanz, V.; and Vetter, T. 2023. A morphable model for the synthesis of 3D faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 157–164.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16123–16133.

Chen, G.; Wang, M.; Yang, Y.; Yu, K.; Yuan, L.; and Yue, Y. 2024. Pointgpt: Auto-regressively generative pre-training from point clouds. *Advances in Neural Information Processing Systems*, 36.

Chung, J. S.; and Zisserman, A. 2017. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, 251–263. Springer.

Fan, Y.; Lin, Z.; Saito, J.; Wang, W.; and Komura, T. 2022. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18770–18780.

Guo, M.-H.; Liu, Z.-N.; Mu, T.-J.; and Hu, S.-M. 2022. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5436–5447.

Guo, Y.; Chen, K.; Liang, S.; Liu, Y.-J.; Bao, H.; and Zhang, J. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5784–5794.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3451–3460.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510.

Ji, X.; Zhou, H.; Wang, K.; Wu, Q.; Wu, W.; Xu, F.; and Cao, X. 2022. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.

Ji, X.; Zhou, H.; Wang, K.; Wu, W.; Loy, C. C.; Cao, X.; and Xu, F. 2021. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14080–14089.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kim, H.; Garrido, P.; Tewari, A.; Xu, W.; Thies, J.; Niessner, M.; Pérez, P.; Richardt, C.; Zollhöfer, M.; and Theobalt, C. 2018. Deep video portraits. *ACM transactions on graphics (TOG)*, 37(4): 1–14.

Xie, Y.; Zhu, J.; Li, S.; and Shi, P. 2023. Cross-modal information-guided network using contrastive learning for point cloud registration. *IEEE Robotics and Automation Letters*, 9(1): 103–110.

Li, J.; Zhang, J.; Bai, X.; Zheng, J.; Ning, X.; Zhou, J.; and Gu, L. 2024. TalkingGaussian: Structure-Persistent 3D Talking Head Synthesis via Gaussian Splatting. *arXiv preprint arXiv:2404.15264*.

Li, J.; Zhang, J.; Bai, X.; Zhou, J.; and Gu, L. 2023. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7568–7578.

Ma, Y.; Zhang, S.; Wang, J.; Wang, X.; Zhang, Y.; and Deng, Z. 2023. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4): 1–15.

Park, K.; Sinha, U.; Hedman, P.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Martin-Brualla, R.; and Seitz, S. M. 2021. HyperNeRF: a higher-dimensional representation for topologically varying neural radiance fields. *ACM Transactions on Graphics (TOG)*, 40(6): 1–12.

Peng, Z.; Hu, W.; Shi, Y.; Zhu, X.; Zhang, X.; He, J.; Liu, H.; and Fan, Z. 2024. SyncTalk: The Devil is in the Synchronization for Talking Head Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peng, Z.; Luo, Y.; Shi, Y.; Xu, H.; Zhu, X.; Liu, H.; He, J.; and Fan, Z. 2023. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5292–5301.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. 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*, 484–492.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30.

Qi, Z.; Dong, R.; Fan, G.; Ge, Z.; Zhang, X.; Ma, K.; and Yi, L. 2023. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning*, 28223–28243. PMLR.

Qian, S.; Kirschstein, T.; Schoneveld, L.; Davoli, D.; Giebenhain, S.; and Nießner, M. 2024. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20299–20309.

Shen, S.; Li, W.; Huang, X.; Zhu, Z.; Zhou, J.; and Lu, J. 2023a. SD-NeRF: Towards Lifelike Talking Head Animation via Spatially-adaptive Dual-driven NeRFs. *IEEE Transactions on Multimedia*.

Shen, S.; Zhao, W.; Meng, Z.; Li, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023b. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1982–1991.

Suwajanakorn, S.; Seitz, S. M.; and Kemelmacher-Shlizerman, I. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4): 1–13.

Tang, J.; Wang, K.; Zhou, H.; Chen, X.; He, D.; Hu, T.; Liu, J.; Zeng, G.; and Wang, J. 2022. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *arXiv preprint arXiv:2211.12368*.

Thies, J.; Elgharib, M.; Tewari, A.; Theobalt, C.; and Nießner, M. 2020. Neural voice puppetry: Audio-driven facial reenactment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, 716–731. Springer.

Wang, J.; Li, X.; Xie, J.; Xu, F.; and Gao, H. 2023a. GaussianHead: Impressive 3D Gaussian-based Head Avatars with Dynamic Hybrid Neural Field. *arXiv e-prints*, arXiv–2312.

Wang, J.; Qian, X.; Zhang, M.; Tan, R. T.; and Li, H. 2023b. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14653–14662.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38(5): 1–12.

Xie, Y.; Zhu, J.; Li, S.; Hu, N.; and Shi, P. 2024. HECPG: Hyperbolic Embedding and Confident Patch-Guided Network for Point Cloud Matching. *IEEE Transactions on Geoscience and Remote Sensing*.

Xing, J.; Xia, M.; Zhang, Y.; Cun, X.; Wang, J.; and Wong, T.-T. 2023. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12780–12790.

Ye, Z.; He, J.; Jiang, Z.; Huang, R.; Huang, J.; Liu, J.; Ren, Y.; Yin, X.; Ma, Z.; and Zhao, Z. 2023. Geneface++: Generalized and stable real-time audio-driven 3d talking face generation. *arXiv preprint arXiv:2305.00787*.

Ye, Z.; Jiang, Z.; Ren, Y.; Liu, J.; He, J.; and Zhao, Z. 2022. GeneFace: Generalized and High-Fidelity Audio-Driven 3D Talking Face Synthesis. In *The Eleventh International Conference on Learning Representations*.

Ye, Z.; Zhong, T.; Ren, Y.; Yang, J.; Li, W.; Huang, J.; Jiang, Z.; He, J.; Huang, R.; Liu, J.; et al. 2024. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. *arXiv preprint arXiv:2401.08503*.

Yu, X.; Tang, L.; Rao, Y.; Huang, T.; Zhou, J.; and Lu, J. 2022. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19313–19322.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2023a. 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*, 8652–8661.

Zhang, Z.; Hu, Z.; Deng, W.; Fan, C.; Lv, T.; and Ding, Y. 2023b. Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3543–3551.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021. 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*, 3661–3670.

Zhong, W.; Fang, C.; Cai, Y.; Wei, P.; Zhao, G.; Lin, L.; and Li, G. 2023. Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.