VividTalk: One-Shot Audio-Driven Talking Head Generation Based on 3D Hybrid Prior

Xusen Sun[1] Longhao Zhang[3] Hao Zhu[1,✉] Peng Zhang[2,✉] Bang Zhang[2]
Xinya Ji[1] Kangneng Zhou[4] Daiheng Gao[2] Liefeng Bo[2]
Xun Cao[1]
[1]Nanjing University [2]Alibaba Group [3]ByteDance [4]Nankai University

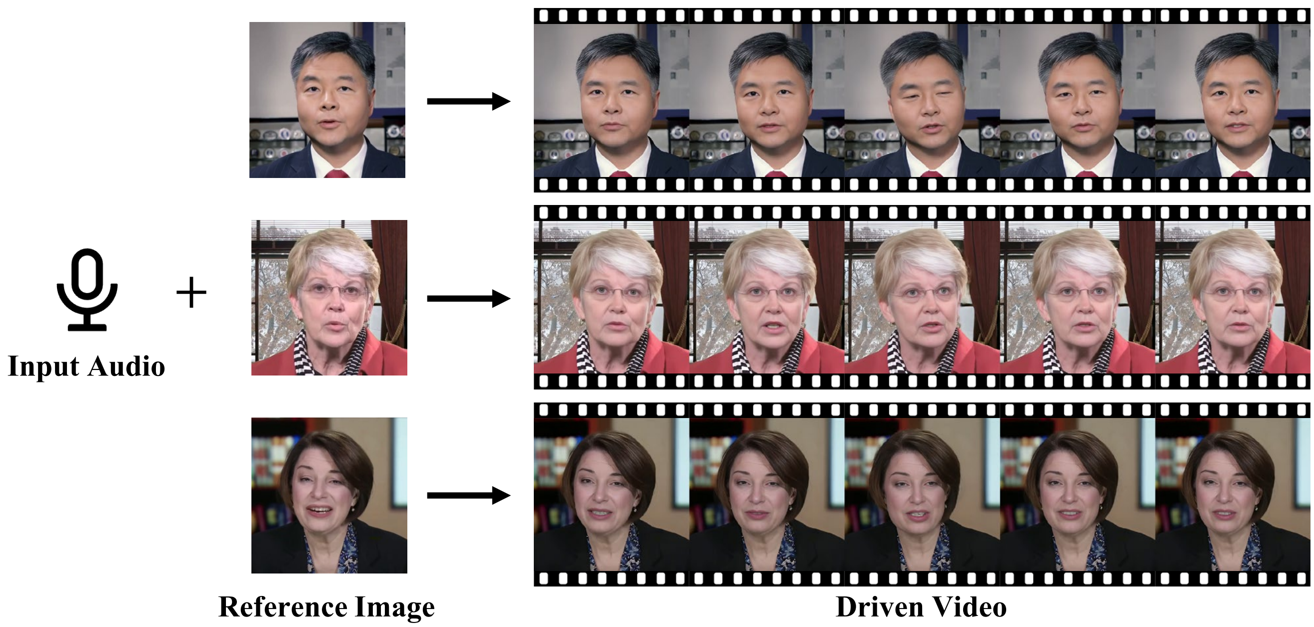

Figure 1. We proposed VividTalk, a generic talking head generation framework. Our method can generate high-visual quality talking head videos with natural facial motion, various head poses, and lip-sync enhanced by a large margin.

Abstract

Audio-driven talking head generation has drawn much attention in recent years, and many efforts have been made in lip-sync, facial motion, head pose generation, and video quality. However, no model has yet led or tied on all these metrics due to the one-to-many mapping between audio and motion. In this paper, we propose VividTalk, a two-stage generic framework that supports generating high-visual quality talking head videos with all the above properties. Specifically, in the first stage, we map the audio to mesh by learning two motions, including non-rigid facial motion and rigid head motion. For facial motion, both blendshape and vertex are adopted as the intermediate representation to maximize the representation ability of the model. For head motion, a novel learnable head pose codebook with a two-phase training mechanism is proposed. In the second stage, we proposed a dual branch motion-vae and a generator to transform the meshes into dense motion and synthesize high-quality video frame-by-frame. Extensive experiments show that the proposed VividTalk can generate high-visual quality talking head videos with lip-sync and realistic enhanced by a large margin, and outperforms previous state-of-the-art works in objective and subjective comparisons. The code will be publicly released upon publication.

1. Introduction

One-shot audio-driven talking head generation aims to drive an arbitrary facial image with speech audio as input signal and has extensive application scenarios, such as virtual avatars [9, 23, 27], visual dubbing [15, 18, 35], and video conferences [5, 33, 37, 39, 43]. As a consequence, it has attracted widespread attention and sparked the interest of numerous researchers. Despite significant efforts invested in researching this topic, state-of-the-art talking heads continue to struggle with achieving realistic and natural performance. We believe that the solution to this problem lies in accurately parsing facial motion and developing a more efficient mapping model that connects speech to animated talking heads.

The overall motion of a talking head mainly comes from two folds: non-rigid facial components and rigid head components. To maximize the photo-realism of the generated videos, both components need to be taken into consideration. For facial motion, most existing approaches adopt a multi-stage framework to map the audio feature to an intermediate representation, *e.g.*, facial landmarks [41, 43], and 3DMM coefficients [38, 40]. However, the facial landmarks are too sparse to model the facial motion in detail. By contrast, the 3D face morphable model [3] (3DMM) has been proven to have the ability to represent the face with high fidelity. Whereas, we observed that the distribution of blendshape on the same facial motion varies considerably, which exacerbates the one-to-many mapping problem between audio and facial motion and leads to a lack of fine-grained motion, especially the lip-related motion. For rigid head motion, it is harder to model because of the weak relationship with audio. Some works [18, 31, 42] utilize a video to provide the head pose or to keep the head still when speaking. Another line of methods [38, 40, 43] present to learn head poses from audio directly, but generate noncontinuous and unnatural results. Up to now, how to generate reasonable head poses from audio is still a challenging problem to be solved.

To address the above problems, we proposed VividTalk, a generic one-shot audio-driven talking head generation framework. Our method only takes a single reference facial image and an audio sequence as inputs, then generates a high-quality talking head video with natural facial motion and various head poses. Specifically, the proposed model is a two-stage framework consisting of Audio-to-Mesh Generation and Mesh-to-Video Generation. In the first stage, considering the one-to-many mapping between facial motion and blendshape distribution, we utilize both blendshape and 3D vertex as the intermediate representation, in which blendshape provides a coarse global motion and vertex offset describes a local fine-grained lip motion. Besides, a multi-branch transformer-based network is also adopted to make full use of long-term audio context to model the relation with the intermediate representations. To learn rigid head motion from audio more reasonably, we cast this problem as a code query task in a discrete and finite space, and build a learnable head pose codebook with a reconstruction and mapping mechanism. After that, both motions learned are applied to reference identity, resulting in driven meshes. In the second stage, based on the driven meshes and reference image, we render the projection texture of both the inner face and outer face, such as the torso, to model the overall motion comprehensively. Then a novel dual branch motion-vae is designed to model the dense motion, which is fed as input to a generator to synthesize the final video in a frame-by-frame manner.

Extensive experiments show that our proposed VividTalk can generate lip-sync talking head videos with natural facial motions and diverse head poses. As shown in Figure 1 and Table 1, both visual results and quantitative analysis demonstrate the superiority of our method in both generated quality and model generalization. To summarize, the main contributions of our work are as follows:

- We present a Global and Local Facial Motion Generator that maps the audio to blendshape and vertex to model natural global facial motion and accurate local lip motion.
- To model rigid head motion more reasonably, a novel Learnable Head Pose Codebook with a two-phase training mechanism is presented.
- A Dual Branch Motion-VAE is proposed to model the dense motion with high accuracy lip movements based on projection texture and 3D landmarks.
- Experiments demonstrate that our proposed VividTalk is superior to the state-of-the-art methods, supporting high-quality talking head video generation and can be generalized across various subjects.

2. Related works

Audio-driven talking head generation. Audio-driven talking head generation aims to drive a facial image according to the audio signal. Early works [4, 7, 30] tried to generate videos in an end-to-end manner. Recently, some works adopted a multi-stage framework to map audio to an intermediate representation, such as 3DMM coefficients [11, 19, 38, 40], facial landmarks [10, 41, 43], and latent code [8, 36] to model the motion better. [38] first generates the 3DMM coefficients from audio, and then the generated 3DMM coefficients are mapped to the unsupervised 3D keypoints to modulate the face render to synthesize videos. [36] introduces a diffusion-baesd framework to generate latent motions from audio condition and noise, and synthesis frames by a decoder. [10] uses facial landmarks and a pre-trained face render to make the generated talking head videos more controllable and high-quality. Similarly, facial landmarks are predicted by [43] to reflect the speaker-aware dynamics to animate both human face images and non-photorealistic cartoon images. [41] only generates lip-related landmarks to inpaint the lower-half occluded facial images. Besides, multiple reference images are needed to produce realistic rendering. However, all of these methods are insufficient to generate lip-sync and realistic talking head videos because of the limitation of the intermediate representation. By contrast, our method uses both blendshape and vertex as the intermediate representation to model the coarse motion and fine-grained motion, respectively.

Video-driven talking head generation. Video-driven talking head generation focuses on transferring the motion of the source actor to the target subject, which is also known as face reenactment. The approaches generally fall into two categories: subject-specific and subject-agnostic. Subject-specific methods [24–26] can produce high-quality videos but can not be extended to new subjects, which limits their application. Recently, some subject-agnostic works [13, 19, 21, 28, 33, 34] have tried to address this problem and achieved tremendous success. For example, [21] disentangles the appearance and motion self-supervised, and learn keypoints along with their local affine transformations to animate the source image. [13] proposes to recover the explicit dense 3D geometry from videos and utilizes the learned depth information to improve the performance of generated talking head videos. Compared to the above methods, our task is more challenging because we need to drive the image with audio as input without any motion prior knowledge.

3. Method

Our method can generate talking head videos with natural facial motion and diverse head poses given an audio sequence and a reference facial image as input. As shown in Figure 2, our framework is composed of two cascaded stages, named Audio-to-Mesh Generation and Mesh-to-Video Generation, respectively. In the following, we first briefly introduce some preliminaries of the 3D morphable model and data preprocessing in Section 3.1. Then, the design of the Audio-to-Mesh stage and Mesh-to-Video stage are described in Section 3.2 and Section 3.3, respectively. Finally, we depicted the training strategy of the total framework in Section 3.4.

3.1. Preliminaries

3D Morphable Model. Our method uses 3D-based (blendshape and vertex) instead of 2D-based information as the intermediate representation for talking head generation. In 3DMM [3], the 3D face shape can be represented as:

$$S = \overline{S} + \alpha U_{id} + \beta U_{exp}, \tag{1}$$

where $\overline{S}$ is the mean shape of the face, U_{id}, and U_{exp} are the PCA bases of identity and expression, respectively. α and β are the identity and expression coefficients for generating a 3D face.

Data Preprocessing. Our model only needs to be trained with an audio-visual synchronized dataset. Before training, some data preprocessing is a prerequisite. Specifically, given a talking head video, we first crop the face region and resize it into 256×256 following [21]. Then the coefficients $\{\alpha \in \mathbb{R}^{150}, \beta \in \mathbb{R}^{52}\}^{\times f}$ and mesh vertices sequence $M^{(3\times n)\times f}$ are reconstructed by [32], where n is the vertex number and f is the frame number. To model the head pose P, rotation matrix $R \in \mathbb{SO}(3)$ and translation vector $t \in \mathbb{R}^3$ are also extracted.

3.2. Audio-to-Mesh Generation

In this section, our goal is to generate 3D-driven meshes according to the input audio sequence and a reference facial image. To be more specific, we first utilize FaceVerse[32] to reconstruct the reference facial image. Next, we learn both non-rigid facial motion and rigid head motion from the audio to drive the reconstructed mesh. To this end, a Global and Local Facial Motion Generator and a Learnable Head Pose Codebook are proposed.

Global and Local Facial Motion Generator. Learning a generic model to generate accurate mouth movements and natural facial motion, *e.g.* eye blinking, with person-specific style is challenging in two aspects: 1) The first challenge is the *audio-motion correlation* problem. As audio signal correlates best with mouth movements, it is difficult to model non-mouth motion from audio. 2) The mapping from audio to facial motion naturally has one-to-many properties, which means that the same audio input may have more than one correct motion pattern, leading to a *mean face* phenomenon with no personal characteristics. To solve the *audio-motion correlation* problem, we use both blendshape and vertex offset as the intermediate representation, for which blendshape provides a coarse facial motion globally and lip-related vertex offset offers a fine-grained lip motion locally. As for the *mean face* problem, we proposed a multi-branch transformer-based generator to model each part's motion individually and inject the subject-specific style to maintain personal features.

Specifically, we utilize a pre-trained audio extractor [1] to extract the contextualized speech representation $A = (a_1, a_2, ..., a_f)$ from the input audio sequence. To represent the person-specific style characteristic, a pre-trained 3D face reconstruction model [32] is used to extract the identity information α from the reference image I_{ref}, which will be encoded as a style embedding z^{style}. Then the audio feature A and the personal style embedding z^{style} are added and fed into a multi-branch transformer-based architecture with two branches to generate blendshape that models facial motion at a coarse level, and the third branch to generate lip-related vertex offset as supplementation of lip motion at a fine-grained level. Note that to model the temporal dependencies better, the learned past motions will be taken as the input of the network when predicting the current motion, which can be formulated as

$$\hat{\beta}_i^f = \Phi_i^{bs}(\hat{\beta}_i^{1...f-1}, A, z^{style}), \quad i \in \{lip, other\}, \tag{2}$$

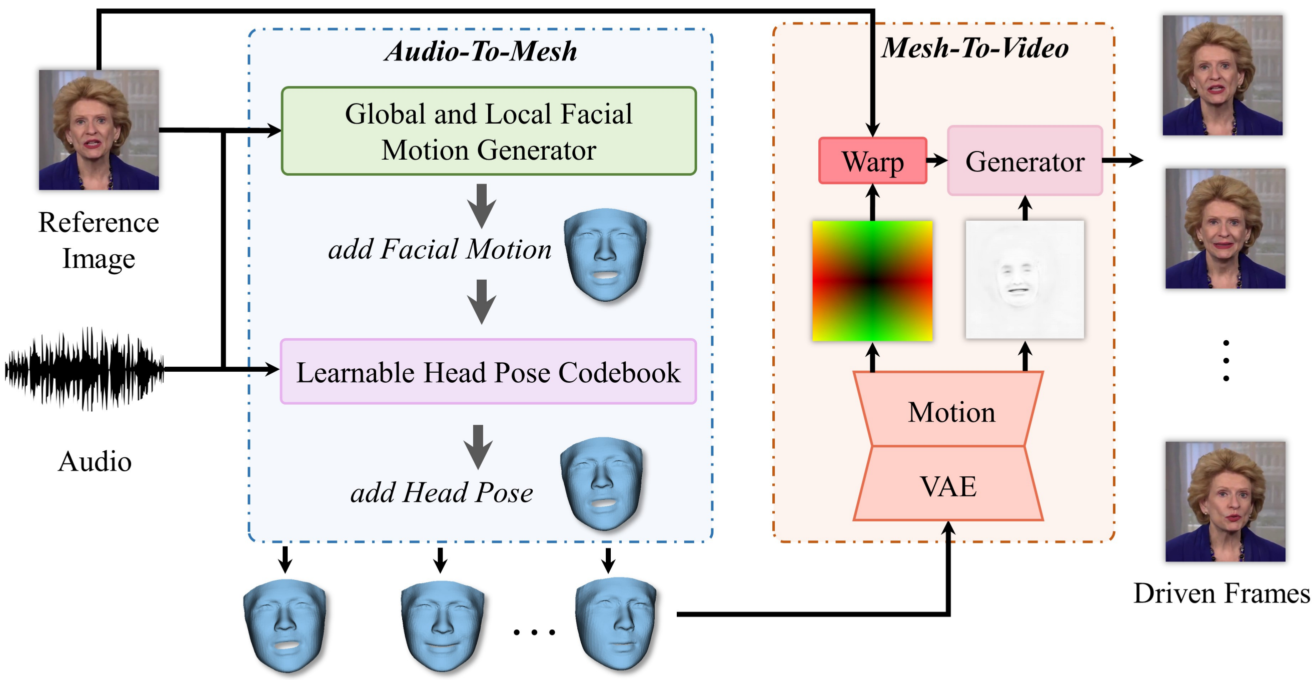

Figure 2. **Overview pipeline.** Our framework consists of two cascaded stages. The Audio-to-Mesh stage maps the audio to non-rigid facial motion and rigid head pose, merged into the facial mesh. The Mesh-to-Video stage transforms the driven meshes into 2D dense motion and synthesizes high-quality and realistic talking head videos.

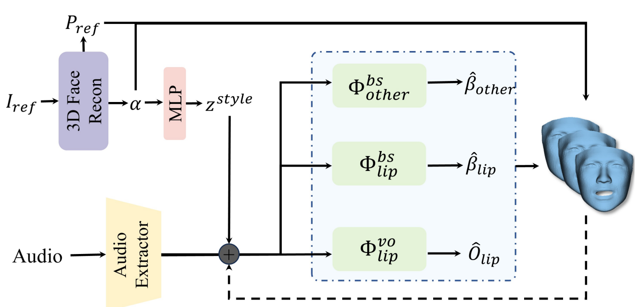

Figure 3. **Global and Local Facial Motion Generator.** The blendshape $\{\hat{\beta}_{lip}, \hat{\beta}_{other}\}$ provide the coarse facial motion with personal style, and the lip-related vertex offset $\hat{O}_{lip}$ supplement the lip motion at a fine-grained level.

$$\hat{O}^{f}_{lip} = \Phi^{vo}_{lip}(\hat{O}^{1\ldots f-1}_{lip}, A, z^{style}), \tag{3}$$

where $\hat{\beta}^{f}_{lip}$, $\hat{\beta}^{f}_{other}$ are the lip-related blendshape and the other blendshape at frame f, respectively. $\hat{O}^{f}_{lip}$ is the lip-related vertex offset at frame f. And Φ is the corresponding network of each branch. Once the training is finished, the driven meshes with non-rigid facial motion can be obtained by

$$\hat{M}_{nr} = (\overline{S} + \alpha U_{id} + (\hat{\beta}_{lip}, \hat{\beta}_{other})U_{exp} + \hat{O}_{lip}) \otimes P_{ref}, \tag{4}$$

where P_{ref} is the pose of reference facial image and $\otimes$ represents the affine transformation caused by P_{ref}.

Learnable Head Pose Codebook. The head pose is another important factor that influences the realism of talking head videos. However, it is not easy to learn it from audio directly because of the weak relationship between them, which will lead to unreasonable and discontinuous results. Inspired by [29] which utilized a discrete codebook as a prior to guarantee high-fidelity generation even with a degraded input. We propose to cast this problem as a code query task in a discrete and finite head pose space and a two-phase training mechanism is carefully designed, with the first phase building an abundant head pose codebook and the second phase mapping the input audio to the codebook to generate the final results, as shown in Figure 4.

In the reconstruction phase, the task is to build a context-rich head pose codebook $\mathcal{Z} = \{z_k\}_{k=1}^{K}$ and a decoder $\mathcal{D}$ with the ability to decode realistic head pose sequence $P^{1:f} \in \mathbb{R}^{6\times f}$ from $\mathcal{Z}$. We adopt a VQ-VAE which constitutes an encoder $\mathcal{E}$, a decoder $\mathcal{D}$, and a codebook $\mathcal{Z}$ as the backbone. Firstly, the relative head pose $P^{1:f}_{r} = P^{1:f} - P^{0}$ is calculated and encoded as a latent code $\hat{Z} = \mathcal{E}(P^{1:f}_{r})$. Then we obtain Z_q using an element-wise quantization function $\mathbf{q}(\cdot)$ to map each item $\hat{z}$ in $\hat{Z}$ to its closest codebook entry z_k:

$$Z_q = \mathbf{q}(\hat{z}) = \underset{z_k \in \mathcal{Z}}{\arg\min} \|\hat{z} - z_k\|\,. \tag{5}$$

Finally, based on the Z_q, the reconstructed relative head pose $\hat{P}^{1:f}_{r}$ is given by the decoder $\mathcal{D}$ as follows:

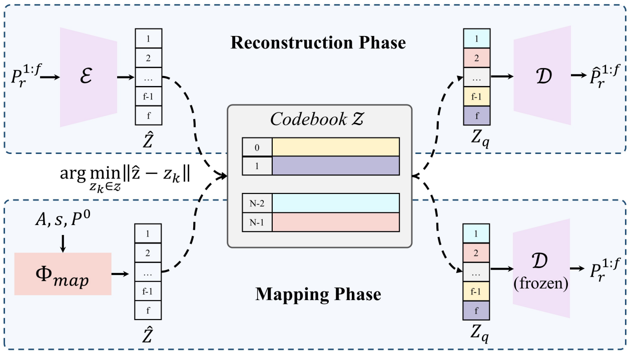

Figure 4. **The two-phase training mechanism of the learnable head pose codebook.** Note that we train the two phases separately, the decoder $\mathcal{D}$ and codebook $\mathcal{Z}$ are frozen during the training of the mapping phase.

$$\hat{P}_r^{1:f} = \mathcal{D}(Z_q) = \mathcal{D}(\mathbf{q}(\mathcal{E}(P_r^{1:f}))). \tag{6}$$

In the mapping phase, we focus on building a network that can map the audio to the codebook learned in the previous phase to generate natural and successive head pose sequences. To model the temporal continuity better, a transformer-based autoregressive model Φ_{map} with self-attention and cross-modal multi-head attention mechanisms was proposed. Specifically, Φ_{map} takes an audio sequence A, person-specific style embedding s and initial head pose P^0 as input, and output an intermediate feature $\hat{Z}$ which will be quantized into Z_q from codebook $\mathcal{Z}$, and then decoded by the pre-trained decoder $\mathcal{D}$:

$$\hat{P}_r^{1:f} = \mathcal{D}(Z_q) = \mathcal{D}(\mathbf{q}(\Phi_{map}(A, s, P^0))). \tag{7}$$

Note that the codebook $\mathcal{Z}$ and the decoder $\mathcal{D}$ are frozen during the training of mapping phase.

So far, both the non-rigid facial motion and rigid head pose have been learned. Now, we can obtain the final driven meshes $\hat{M}_d$ by applying the learned rigid head pose to mesh $\hat{M}_{nr}$:

$$\hat{M}_d^{1:f} = \hat{M}_{nr}^{1:f} \otimes \hat{P}_r^{1:f}. \tag{8}$$

3.3. Mesh-to-Video Generation

This section is devoted to transforming the driven meshes into videos. As shown in Figure 5, a dual branch motion-VAE is proposed to model the 2D dense motion, which will be taken as the input of the generator to synthesize the final video. Next, we will introduce this process in detail.

Transforming 3D domain motion to 2D domain motion directly with a neural network is difficult and inefficient because the network needs to seek the correspondence between two domain motions for better modeling. To decrease the learning burden of the network and achieve further performance, we conduct this transformation in the 2D domain with the help of projection texture representation.

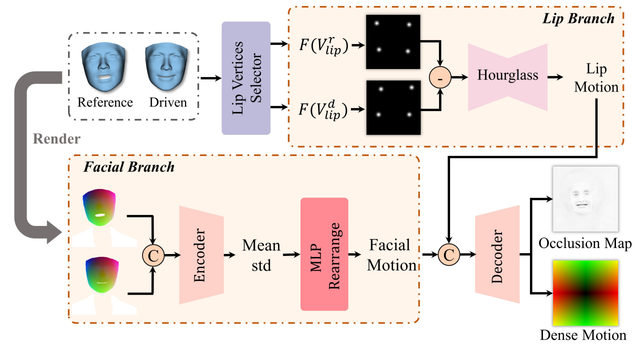

Figure 5. **Dual branch motion-VAE.** The lower branch models the global facial motion; The upper branch augments the lip motion based on lip-related landmarks.

To render the projection texture of 3D mesh, we first normalize the mean shape of the 3D face to $0-1$ in x, y, z axis to obtain a Normalized Coordinate Code NCC with three channels similar to RGB, which can be seen as a new representation of the face texture:

$$NCC_i = \frac{\overline{S}_i - \min(\overline{S}_i)}{\max(\overline{S}_i) - \min(\overline{S}_i)}, \quad i \in \{x, y, z\}. \tag{9}$$

Then we adopt Z-Buffer to render the projected 3D inner face texture PT_{in} colored by NCC. However, the outer face region can not be modeled well because of the limitation of 3DMM. To model the motion across frames better, we use [16] to parse images and obtain the outer face region texture PT_{out}, such as the torso and background, which will be combined with PT_{in} as below:

$$PT = PT_{in} \cdot M + PT_{out} \cdot (1 - M), \tag{10}$$

where M is the mask of the inner face.

As shown in Figure 5, in the facial branch, the reference projected texture PT_{ref} and driven projected texture PT_d are concatenated and fed into an Encoder followed by an MLP, which outputs a 2D facial motion map. To further enhance lip movements and model more accurately, we also select lip-related landmarks and transform them into Gaussian maps, a more compact and efficient representation. Then an Hourglass network takes the substracted Gaussian map as input and outputs a 2D lip motion, which will be concatenated with the facial motion and decoded into a dense motion and an occlusion map.

Finally, we warp the reference image based on the dense motion map predicted before and obtain the deformed image, which will be taken as the input to the generator with the occlusion map to synthesize the final video frame by frame.

3.4. Training Strategy

Training such a framework is not easy. Specifically, we train the Audio-to-Mesh stage and Mesh-to-Video stage separately. And the complete framework can be inferred in an end-to-end fashion. The Global and Local Facial Motion Generator are supervised by reconstruction loss in terms of blendshape and mesh:

$$L_{bsvo} = \left\| \beta - \hat{\beta} \right\| + \left\| M - \hat{M}_{nr} \right\|. \tag{11}$$

In the training of Learnable Head Pose Codebook, due to the quantization function (5) is not differentiable, we apply a straight-through gradient estimator [2] that copies the gradients from the decoder to the encoder. Then the two-phase training is supervised as follows:

$$\begin{aligned} L_{rec} = & \left\| P_r^{1:f} - \hat{P}_r^{1:f} \right\|^2 + \left\| sg(\mathcal{E}(P_r^{1:f})) - z_q \right\|_2^2 \\ & + \left\| sg(z_q) - \mathcal{E}(P_r^{1:f}) \right\|_2^2, \end{aligned} \tag{12}$$

$$L_{map} = \left\| P_r^{1:f} - \hat{P}_r^{1:f} \right\|^2 + \left\| \hat{Z} - sg(Z_q) \right\|_2^2, \tag{13}$$

where $sg(\cdot)$ denotes a stop-gradient operation.

As for the Mesh-to-Video stage, the perceptual loss L_{perc} based on the pre-trained VGG-19 [22] network is used as the main driving loss. The feature matching loss L_{fm} is also used to stabilize the training as the generator has to produce realistic results.

4. Experiments

4.1. Dataset and Metrics

Dataset. We train our model with the HDTF [40] dataset and VoxCeleb [17] dataset. HDTF is a high-resolution audio-visual dataset containing over 16 hours of video on 346 subjects. VoxCeleb is another larger dataset involving more than 100k videos and 1000 identities. We first filter the two datasets to remove the invalid data, *e.g.*, data with out-of-sync audio and video. Then following the [21], we leverage a face landmarks detector to crop the face region in the video and resize them into 256×256. Finally, the processed videos are divided into $80\%, 10\%, 10\%$, which will be used for training, validating, and testing.

Metrics. To demonstrate the superiority of the proposed method, we evaluate the model with several metrics. The SyncNet score [6] is utilized to measure lip synchronization quality, which is an important indicator for talking head applications. To evaluate the realism and identity preservation of the results, we calculate the Frechet Inception Distance (FID) [12] and cosine similarity (CSIM) between the reference image and generated frames. Besides, the variance of the norm of head pose in the Euler angle representation is calculated to evaluate head pose diversity (HPD) better.

4.2. Implementation Details

In our experiments, we use FaceVerse [32], the state-of-the-art single image reconstruction method to recover the video and obtain the ground truth blendshapes and meshes for supervision. During training, the Audio-to-Mesh stage and Mesh-to-Video stage are trained separately. Specifically, the Global and Local Facial Motion Generator and Learnable Head Pose Codebook in the Audio-to-Mesh stage are also trained separately. During inference, our model can work in an end-to-end manner by cascading the above two stages. For optimization, the Adam optimizer [14] is used with the learning rate 1×10^{-4} and 1×10^{-5} for two stage, respectively. And the total training costs 2 days on 8 NVIDIA V100 GPUs. More details about training and network architecture can be referred to in the supplementary material.

4.3. Comparison with state-of-the-art methods

We qualitatively and quantitatively compare the proposed method to several prior state-of-the-art works on audio-driven talking head generation, including the SadTalker [38], TalkLip [31], MakeItTalk [43], Wav2Lip [18], and PC-AVS [42]. The experiments are conducted in Same-Identity Reconstruction and Cross-Identity Dubbing setting. In the Same-Identity Reconstruction setting, the audio signal and the reference image come from the same identity. While in the Cross-Identity Dubbing setting, videos non-existent in the world are generated because the audio comes from another person.

Qualitative Comparison. Figure 6 demonstrates the visual results of our method and previous methods. It can be seen that SadTalker [38] fails to generate accurate fine-grained lip motion and is inferior to our video quality. This is because it only uses the blendshape as the intermediate representation which is insufficient to model the accurate lip motion. TalkLip [31] generates blurry results and changes the skin color style to slightly yellow, which loses the identity information to a certain degree. MakeItTalk [43] can not generate accurate mouth shapes, especially in the Cross-Identity Dubbing setting. Wav2Lip [18] tends to synthesize blurry mouth regions, and output video with static head pose and eye movement when inputting a single reference image. PC-AVS [42] requires a driven video as input and struggles for identity preservation. By contrast, our proposed method can generate high-quality talking head videos with accurate lip-synchronized and natural facial motion.

Quantitative Comparison. As shown in Table 1, our method performs better in image quality and identity preservation, which is reflected by lower FID and higher CSIM metrics. Thanks to the novel learnable codebook mechanism, the head pose generated by our method is also more diverse and natural. Though the SyncNet score of our method is inferior to Wav2Lip [18], our method can drive the reference image with single audio instead of video and generate frames in higher quality.

Method	Head Pose Generation	Same-Identity Reconstruction				Cross-Identity Dubbing		
		SyncNet ↑	FID ↓	CSIM ↑	HPD ↑	SyncNet ↑	CSIM ↑	HPD ↑
Real Video	✗	7.838	0.000	1.000	0.217	✗	✗	✗
SadTalker [38]	✓	5.711	28.35	0.862	0.305	5.416	0.849	0.337
TalkLip [31]	✗	5.503	23.18	0.713	✗	5.295	0.686	✗
MakeItTalk [43]	✓	3.346	33.73	0.845	0.286	3.128	0.840	0.291
Wav2Lip [18]	✗	**6.757**	21.80	0.816	✗	**6.127**	0.807	✗
PC-AVS [42]	✗	6.404	84.67	0.674	✗	5.538	0.613	✗
Ours	✓	6.684	**20.32**	**0.916**	**0.437**	6.018	**0.907**	**0.497**

Table 1. **Quantitative comparison.** We compare our results with state-of-the-art talking head generation works. Note that our proposed VividTalk outperforms previous works in lip-sync, video quality, identity preservation, and head pose diversity.

Figure 6. **Qualitative comparison.** SadTalker [38] and MakeItTalk [43] can generate results with a single image and audio as input. While TalkLip [31], Wav2Lip [18], and PC-AVS [42] need another video to provide the head poses for the final results. We highly recommend watching our supplementary video for a more intuitive comparison.

4.4. User Study

To further evaluate the proposed method, we conducted a user study with 20 volunteers to rate the videos generated by each method. For a fair comparison, 10 in-the-wild facial images with various characteristics and poses are selected as reference images, and 5 audio with diverse languages and speaking styles are chosen as driven signals, which are taken as the input of each method and generate 50 videos in total. The volunteers are invited to rate each video between 1 and 5 (higher is better) in terms of lip synchronization, motion naturalness(both facial motion and

Method	Lip Sync	Motion Naturalness	Identity Preservation	Overall Quality
SadTalker [38]	3.891	3.107	4.035	3.626
TalkLip [31]	3.217	✗	3.891	3.418
MakeItTalk [43]	2.836	2.748	3.740	2.914
Wav2Lip [18]	2.751	✗	3.814	2.471
PC-AVS [42]	3.106	✗	2.603	2.513
Ours	**4.315**	**3.896**	**4.618**	**4.307**

Table 2. User study.

head motion), identity preservation, and overall quality. As shown in Table 2, the final mean score of our method outperforms previous methods in all metrics, which indicates the superiority of our method.

4.5. Ablation Study

In this section, we conduct several ablation studies to verify the effectiveness of each design in proposed method.

Figure 7. Ablation about Intermediate representation.

Ablation about Intermediate Representation. To verify the superiority of using both blendshape and vertex offset as the intermediate representation, we implement two variant models using either blendshape or vertex offset to generate facial meshes from audio. The final driven results are shown in Figure 7. We can see that the method using blendshape only as the intermediate representation can model most facial motion well but not lip motion. The method using vertex offset as the intermediate representation can model the mouth shape better but lead to artifacts in the teeth region. By comparison, the method with both representations can generate accurate and fine-grained motion with high video quality maintained.

Ablation about Learnable Head Pose Codebook. We also perform experiments to validate the design effectiveness of the Learnable Head Pose Codebook. On the one hand, we learn absolute instead of relative head pose from the audio. On the other hand, we remove the initial head pose P^0 as a condition in the mapping phase. As shown in Table 3, learning absolute head pose leads to lower diversity, and our method without the initial head pose results in an unnatural visual effect. By contrast, our full method performs better in both evaluation metrics, indicating the benefits of our designs.

Ablation about dual branch motion-VAE. We evaluate the proposed dual branch motion-VAE in Mesh-to-Video stage regarding lip synchronization and video quality. Specifically, we designed a variant that keeps the facial motion branch only and removes the lip motion branch. As shown in Figure 8, the method without the lip-motion branch can not model the mouth shape accurately and generates frames with artifacts in the teeth area. In contrast, the dual branch model can synthesize results well benefits from the enhancement of the lip-motion by lip-branch.

Method	Diversity ↑	Naturalness ↑
Absolute Head Pose Prediction	0.379	3.641
w/o Initial Head Pose	0.408	3.728
Our Full	**0.437**	**3.896**

Table 3. Ablation about Learnable Head Pose Codebook.

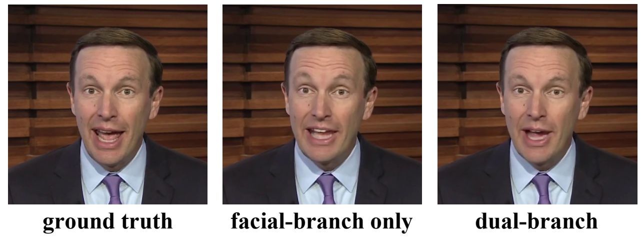

Figure 8. Ablation about dual branch motion-VAE.

5. Conclusion

In this paper, we proposed VividTalk, a novel and generic framework supporting the generation of high-quality talking head videos with natural facial motions and diverse head poses. For non-rigid facial motion, both blendshape and vertex are mapped as the intermediate representation to maximize the representation of the model, and an elaborate multi-branch generator is designed to model global and local motion individually. As for rigid head motion, a novel Learnable Head Pose Codebook with a two-phase training mechanism is proposed to synthesize natural results. Thanks to the dual branch motion-VAE and generator, the driven meshes can be transformed into dense motion well and used to synthesize finale videos. Experiments demonstrate our method outperforms previous state-of-the-art methods and opens new avenues in many applications, such as digit human creation, video conferences, and so on.

Limitation. As it stands, VividTalk does not account for emotional nuances, thereby hindering its effectiveness in creating talking head videos with expressive emotional content. A further challenge arises when the model attempts to generate realistic teeth from reference images where the mouth is closed, potentially leading to unsatisfactory results. We consider addressing the issue with the inpainting power of diffusion models [20]. Specifically, one approach worth attempting is to inpaint the teeth of the first frame and generate subsequent frames based on the first frame. We will explore to solve these issues in the future work.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020. 3

[2] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 6

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

[4] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision*, pages 520–535, 2018. 2

[5] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *Proceedings of the European Conference on Computer Vision*, pages 35–51. Springer, 2020. 1

[6] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Proceedings of the Asian Conference on Computer Vision Workshops*, pages 251–263. Springer, 2017. 6

[7] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017. 2

[8] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8498–8507, 2024. 2

[9] Kuangxiao Gu, Yuqian Zhou, and Thomas Huang. Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10861–10868, 2020. 1

[10] Siddharth Gururani, Arun Mallya, Ting-Chun Wang, Rafael Valle, and Ming-Yu Liu. Spacex: Speech-driven portrait animation with controllable expression. *arXiv preprint arXiv:2211.09809*, 2022. 2

[11] Yuxiao He, Yiyu Zhuang, Yanwen Wang, Yao Yao, Siyu Zhu, Xiaoyu Li, Qi Zhang, Xun Cao, and Hao Zhu. Head360: Learning a parametric 3d full-head for free-view synthesis in 360 {\deg}. *arXiv preprint arXiv:2408.00296*, 2024. 2

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 6

[13] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3397–3406, 2022. 3

[14] D Kinga, Jimmy Ba Adam, et al. A method for stochastic optimization. In *International Conference on Learning Representations*, page 6. San Diego, California;, 2015. 6

[15] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. Towards automatic face-to-face translation. In *Proceedings of the ACM International Conference on Multimedia*, pages 1428–1436, 2019. 1

[16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 5

[17] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 6

[18] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the ACM International Conference on Multimedia*, pages 484–492, 2020. 1, 2, 6, 7, 8

[19] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021. 2, 3

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 8

[21] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019. 3, 6

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[23] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*, 2018. 1

[24] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics*, 36 (4):1–13, 2017. 3

[25] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Transactions on Graphics*, 34(6):183–1, 2015.

[26] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016. 3

[27] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *Proceedings of the European Conference on Computer Vision*, pages 716–731. Springer, 2020. 1

[28] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Facegan: Facial attribute controllable reenactment gan. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1329–1338, 2021. 3
[29] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017. 4
[30] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128:1398–1413, 2020. 2
[31] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. 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*, pages 14653–14662, 2023. 2, 6, 7, 8
[32] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20333–20342, 2022. 3, 6
[33] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2021. 1, 3
[34] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. Anifacegan: Animatable 3d-aware face image generation for video avatars. *Advances in Neural Information Processing Systems*, 35:36188–36201, 2022. 3
[35] Tianyi Xie, Liucheng Liao, Cheng Bi, Benlai Tang, Xiang Yin, Jianfei Yang, Mingjie Wang, Jiali Yao, Yang Zhang, and Zejun Ma. Towards realistic visual dubbing with heterogeneous sources. In *Proceedings of the ACM International Conference on Multimedia*, pages 1739–1747, 2021. 1
[36] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667*, 2024. 2
[37] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9459–9468, 2019. 1
[38] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023. 2, 6, 7, 8
[39] Xi Zhang, Xiaolin Wu, Xinliang Zhai, Xianye Ben, and Chengjie Tu. Davd-net: Deep audio-aided video decompression of talking heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12335–12344, 2020. 1
[40] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 2, 6
[41] Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2023. 2
[42] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4186, 2021. 2, 6, 7, 8
[43] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics*, 39(6):1–15, 2020. 1, 2, 6, 7, 8