Speech2Lip: High-fidelity Speech to Lip Generation by Learning from a Short Video

Xiuzhe Wu¹, Pengfei Hu², Yang Wu³, Xiao yang Lyu¹, Yan-Pei Cao³, Ying Shan³, Wenming Yang², Zhongqian Sun⁴, Xiaojuan Qi¹,

¹The University of Hong Kong, ² Tsinghua University, ³ ARC Lab, Tencent PCG, ⁴ Tencent AI Lab

Abstract

Synthesizing realistic videos according to a given speech is still an open challenge. Previous works have been plagued by issues such as inaccurate lip shape generation and poor image quality. The key reason is that only motions and appearances on limited facial areas (e.g., lip area) are mainly driven by the input speech. Therefore, directly learning a mapping function from speech to the entire head image is prone to ambiguity, particularly when using a short video for training. We thus propose a decomposition-synthesis-composition framework named Speech to Lip (Speech2Lip) that disentangles speech-sensitive and speech-insensitive motion/appearance to facilitate effective learning from limited training data, resulting in the generation of natural-looking videos. First, given a fixed head pose (i.e., canonical space), we present a speech-driven implicit model for lip image generation which concentrates on learning speech-sensitive motion and appearance. Next, to model the major speech-insensitive motion (i.e., head movement), we introduce a geometry-aware mutual explicit mapping (GAMEM) module that establishes geometric mappings between different head poses. This allows us to paste generated lip images at the canonical space onto head images with arbitrary poses and synthesize talking videos with natural head movements. In addition, a Blend-Net and a contrastive sync loss are introduced to enhance the overall synthesis performance. Quantitative and qualitative results on three benchmarks demonstrate that our model can be trained by a video of just a few minutes in length and achieve state-of-the-art performance in both visual quality and speech-visual synchronization. Code: https://github.com/CVMI-Lab/Speech2Lip.

1. Introduction

Learning from a short video to generate personalized audio-synchronized talking videos driven by a speech is of great importance to various applications, for instance, digital human animation, video dubbing, and UGC video creation. However, synthesizing high-fidelity videos from speech for a desired speaker remains a challenging task.

The first challenge arises from complicated motion patterns. Although speech majorly influences lip areas (i.e., speech-sensitive areas), lip movements are often accompanied by other motions, such as global head movements, which greatly impact lip shapes and appearances. Thus, directly synthesizing a whole image from speech often leads to inaccurate lip synthesis. Second, existing methods still struggle to satisfy appearance fidelity requirements, which include both identity preservation (speaker-specific) and high-quality detail generation, such as clear details of teeth, tongue, and eye-blinking [9, 42, 26, 44, 7]. Third, it’s difficult to acquire videos longer than 10 hours for a speaker which is yet required by conventional methods [17, 31].
To tackle the aforementioned challenges, existing attempts can be coarsely categorized into two lines of research: speaker-independent and speaker-specific methods. The first line often exploits GANs [15] that need to be trained on large-scale multi-person datasets. However, GAN-based models [4, 7, 9, 26, 33, 38, 42, 43] usually synthesize low-resolution images and unnatural motions (i.e., background movements). They thus hardly meet the appearance fidelity requirement in terms of sharpness and fine appearance details. Moreover, preserving the identity of the speaker remains a challenging task [7, 43].

For the consideration of high-fidelity and identity preservation, another strategy focuses on learning from a specific speaker. Although attaining high-fidelity results, early works [17, 31] often require several hours of video footage from a speaker for training, hindering their practical applicability. Recently, NeRF [20] has emerged as a promising approach for generating high-fidelity videos, which succeeds in learning from a short video of just a few minutes and having the potential to generate high-fidelity results [16, 18, 28]. Nevertheless, the models still struggle with appearance and motion ambiguity issues because they model speech-sensitive motions/appearances together with other facial areas less correlated to the given speech. This issue becomes more severe when training data is limited since no extra information can be exploited to avoid interference from signals that are not correlated with the speech. As a result, they tend to generate lip sequences that do not synchronize well with the speech and produce blurry images (see Figure 5 and Table 1). Therefore, reducing the complexity of modeling motions is critical to enable effective learning from a short video and synthesizing high-quality videos for a specific speaker.

We thus design a preliminary experiment to identify that motion and appearance of lip areas have a strong correlation with speech, while head motion and other facial areas are less related to speech (Figure 2). Motivated by the observation, we propose decomposing speech-insensitive motion/appearance from speech-sensitive one, and synthesizing them separately before composing them into a new talking video that aligns with the given speech (Figure 1). Toward this goal, we present a decomposition-synthesis-composition framework named Speech to Lip (Speech2Lip). In the decomposition stage, we introduce a speech-driven implicit model that generates high-fidelity synced lip sequences in a fixed view (i.e., canonical view). To model 3D head motion effectively, we design a Geometry-Aware Mutual Explicit Mapping (GAMEM) module that estimates explicit geometric mappings between an arbitrary observed view and the canonical view. GAMEM also includes a jointly optimized canonical-view full-head depth map, which enables the model to be 3D-aware and supports controllable synthesis driven by new head poses. In the composition stage, GAMEM allows us to flexibly paste the synthesized canonical-view lips onto an arbitrary observed view to obtain natural synchronized talking videos. To improve the synthesis and synchronization qualities after composition, we incorporate a blending network (i.e., Blend-Net) to refine the results and a contrastive sync loss to facilitate learning from a short video for generating synchronized talking videos.

Our major contributions are summarized as follows:

1) We introduce a novel framework that disentangles speech-sensitive and speech-insensitive motion/appearance in high-fidelity video synthesis. By separating these components, the framework can effectively learn from limited training data.

2) The proposed speech-driven implicit model synthesizes speech-sensitive contents and the GAMEM flexibly combines them with given speech-insensitive areas to generate synchronized talking heads with natural movements and support pose-controllable synthesis.

3) Both qualitative and quantitative experimental results demonstrate the superiority of our method over the state-of-the-art speaker-specific methods.

2. Related Work

Speech-Driven Talking Face Synthesis. Video synthesis from speech is a long-standing problem. Recent works can be generally divided into two categories: speaker-specific [17, 31, 19, 16, 39, 18, 28] and speaker-independent [9, 42, 26, 44, 7, 5, 41, 35, 36]. In the first track, earlier works have succeeded in obtaining realistic visual results for a target person [17, 31] but required hours of video belonging to one specific speaker. Therefore, many efforts [19, 16, 39, 18, 28] have been made to train the model with a shorter talking video (3-5 minutes). However, they utilize speech to forecast overall motions including speech-insensitive ones, failing to learn accurate lip shapes and appearances. The other line (i.e., speaker-independent method) aims to build a universal model for all identities [9, 42, 26, 44, 7]. The end-to-end pipeline [9, 42] is usually developed on GAN [15]. To boost performance, Richard et al. [27] only synthesize the mesh of limited facial areas, Prajwal et al. [26] present an extra lip-sync discriminator, Zhou et al. [43] provide additional head poses as inputs, and some other models [7, 14, 44, 32, 30] leverage the intermediate representation (e.g., 2D facial landmarks [7, 14, 44] or expression parameters [32, 30]). Nevertheless, they often suffer from low video quality, difficult identity preservation, and abnormal head motion generation.

Implicit Representation based Talking Head Methods. Recently, implicit representations have shown high model-
ing capabilities in multiple tasks [25, 20, 24, 34]. Among them, NeRF [20] obtains extraordinary performance in novel view synthesis by training on only hundreds of images. Guo et al. [16] first apply it in the speaker-specific talking head synthesis task to enable learning from a short video. However, they utilize two NeRFs to model the head and torso, resulting in the head-torso separation phenomenon. Liu et al. [18] solve it by leveraging one unified NeRF with two semantic-aware modules. Shen et al. [28] incorporate 2D image features as additional inputs to further reduce training data requirements (i.e., videos with only 10-15 seconds). Regardless, these methods only ensure visual quality, and simply use speech to drive all the complex motions, leading to the out-of-sync problem.

3. Empirical Study and Motivations

Our approach is motivated by the observation that only limited facial areas are highly correlated to speech. To verify this, we conduct a preliminary experiment to determine which areas are most speech-sensitive. Specifically, we apply warping to all captured images with varying head poses (refer to observed views) using the 3D Morphable Model (3DMM) [2] and bring them to a fixed head pose (known as the canonical view). We then compute a motion heatmap in the canonical view to determine the areas most responsive to speech. As depicted in Figure 2, the elimination of head motion results in most of the motions around the lip regions displaying high sensitivity to the input speech. Additional examples can be found in the supplementary file. For more details on the warping process, please refer to Sec. 4.3.

4. Method

4.1. Overview

Figure 3 provides an overview of our model. We define each frame captured in the observed view with its own head motion as the “observed space”. To disentangle speech-sensitive and speech-insensitive motions (i.e., head motions), we randomly select one observed space to serve as the “canonical space”. By doing so, we are able to align all observed spaces with the canonical space, thereby eliminating the effects of head motions. Note that once selected, the canonical space is fixed during training and inference. To model speech-sensitive motions, we propose a speech-driven implicit model to generate lip images in the canonical space without head motion effect (Sec. 4.2). Thus, the generated lip image is canonical-view. A Geometry-Aware Mutual Explicit Mapping (GAME-M) module is further proposed to model the speech-insensitive head motions without speech effect. Next, we project the lip images in canonical space generated by the above implicit model to the observed space based on GAME-M so that the synthesized lip image can be aligned and composed with any arbitrary observed frame, giving the model the flexibility for diversified composition (Sec. 4.3). Finally, a blending network (i.e., Blend-Net) and a contrastive sync loss are presented to improve the quality of final synthetic images $I_o$ (Sec. 4.4).

4.2. Disentangled and Synced Implicit Modeling

The central part of Speech2Lip is the disentangled and synced speech-driven implicit generator (Figure 3(a)), which focuses on generating 2D lip appearances that are synchronized with cross-modal speech. By employing motion and appearance disentanglement, the generator only retains speech-sensitive motions and appearances. A lip-syncing contrastive loss is utilized (Figure 3(c)) so that it can generate synced high-fidelity mouth appearance by learning from very short video data. All these together make the implicit modeling disentangled and synced.

Speech-driven Implicit Model To achieve high-fidelity canonical lip image generation, we utilize a speech-driven implicit model. An implicit model is defined as a function that maps the coordinate signal to another signal [29], e.g., color. The input coordinates of the implicit model are defined in a continuous space (i.e., the real field) so that it helps exploit the inherent relationship between spatially adjacent locations. Also, it can be further mapped into the frequency domain and benefit the learning of high-frequency information [20]. Besides, by learning a dedicated model for a specific scenario, implicit models also gain a strong ability to generate high-quality desired outputs. In our model, color information $\hat{c}_{c,n} = (r, g, b)$ is regarded as our output signal (appearance color), and the canonical implicit function $f_\theta$ can be defined as

$$\hat{c}_{c,n} = f_\theta(x_{c,n}, \alpha, t_s),$$  

where $x_{c,n}$ is the continuous 2D pixel coordinate vector $(u,v)$ in the canonical space. $\alpha \in \mathbb{R}^{64}$ is the feature vector of the input speech at the concerned moment and $t_s$ represents the timestamp information, which is utilized to enhance the temporal consistency between adjacent frames.

Continuous Sampling The next question that arises is how to obtain the corresponding supervision signal to train the implicit model, which poses a challenge as the coordinates of the pixels are always integers, while we define the coordinates in a continuous space for more expressive features. To overcome this challenge, we adopt the approach developed in [8], which enables us to sample and generate corresponding supervisions in the continuous coordinate space.
Particularly, for each pixel $x_c$, a rectangle with an arbitrary shape is randomly sampled, the four corner points (i.e., $x_{c,n}$, $n \in \{00, 01, 10, 11\}$) of which are regarded as our inputs. A weighted averaged value $\hat{c}_c$ is then calculated as

$$\hat{c}_c = \sum_{n \in \{00, 01, 10, 11\}} \frac{S_n}{S} \cdot \hat{c}_{c,n},$$

where $\hat{c}_{c,n}$ are the output appearance values of the sampled points. As illustrated in Figure 4, the blue-dashed-line rectangle represents the area $S$, which is divided into four sub-rectangles by gray lines, each with an area $S_n$. By leveraging this strategy, we can generate supervision for points with continuous positions, which enables us to fully exploit the advantages of the implicit model and achieve high-fidelity visual results.

### 4.3. Geometry-Aware Mutual Explicit Mapping

Once we have obtained the speech-sensitive lip sequences in the canonical view, we can construct the eventually synthesized image $I_o$ by assigning the synthesized pixels to a location in the observed space. However, since the assigning process is speech-insensitive and only relies on the geometry information, we introduce the GAMEM module, illustrated in Figure 3, to explicitly model it. The GAMEM module comprises Canonical-to-Observed (C2O), Observed-to-Canonical (O2C) Mapping, and a learnable full-head depth map, which supports various applications, including pose-controllable image synthesis.

**C2O/O2C Mapping** These mappings represent how the images transform between two spaces (i.e., canonical space and observed space). As depicted in Figure 3, O2C Mapping aims at creating supervisions $I_{o\rightarrow c}$ for training the lip generator model $f_o$ and C2O Mapping is utilized to warp the generated lip image into the observed view to produce $I_{c\rightarrow o}$ for further composition. As they follow similar principles, we take the O2C Mapping as an example to illustrate the process. The input of O2C Mapping contains $I_o$ and $I_c$, and the output are pixel correspondences that map pixel coordinates in observed space to the canonical space. To achieve it, we use the 3D Morphable Model (3DMM) [2] to calculate the overall camera intrinsic matrix $K$ and rough face geometry $G$. The head poses $T_o \in \mathbb{R}^{4 \times 4}$ and $T_c \in \mathbb{R}^{4 \times 4}$, which include a rotation matrix and a translation vector, are also estimated. And the relative head pose $T_{c\rightarrow o} \in \mathbb{R}^{4 \times 4}$ between two spaces can be computed as

$$T_{c\rightarrow o} = T_o \times T_c^{-1}.$$  

Furthermore, face depth maps $D_o$ and $D_c$ for $I_o$ and $I_c$ obtained from 3DMM can be further interpolated based on $G$, and the corresponding head poses. Noticing that 3DMM can only be aware of the face area instead of the head area,
but the information about the rest part of the head (e.g., forehead, ears, eyes, mouth) is lacking. Hence we optimize the \( D_c \), along with our model to complete the missing depth region. If \( p_o \) and \( p_c \) are the corresponding 2D homogeneous pixel grid coordinates on \( I_o \) and \( I_c \), the geometric relationships can be explicitly formulated as

\[
D_o p_o = K T_{c \rightarrow o} D_c K^{-1} p_c, \tag{4}
\]

so the position mapping from canonical space to observed space can be easily obtained as

\[
F_{c \rightarrow o}(T_{c \rightarrow o}, D_c, p_c) = p_o. \tag{5}
\]

Then, we can warp the observed ground-truth lip images \( I^m_o \) into the canonical space by backward warping, denoted by \( I^m_o \rightarrow c \), to guide the learning of the lip generation model. The C2O Mapping can be similarly defined by \( T_{o \rightarrow c} \) and \( D_o \). With it, we can warp the generated canonical lip image \( I^m_c \) into the observed space by backward warping, denoted by \( I^m_c \rightarrow o \), to compose with the given observed image \( I_o \) for synthesizing a talking face.

**Pose Controllable Synthesis** Thanks to the simultaneously learned complete depth map \( D_o \), pose-controllable full-head image synthesis according to users’ requirements is also supported in our model. Specifically, we employ \( D_c \) to determine location correspondences based on Eq. 3 and Eq. 5 by altering \( T_o \). There is a slight difference in this situation, which is to use the forward warping strategy as the ground-truth depths at new given poses are missing, which will produce black holes. To mitigate this issue, we propose a data augmentation method in training by randomly adding black holes with a ratio of 50% to the image to let the Blend-Net learn to fill these areas.

### 4.4. Overall Refinement

**Image Blending** With \( I^m_{c \rightarrow o} \), we can integrate it by directly pasting it to the original observed frame based on 2D lip landmarks [3]. However, the paste operation often results in mismatching artifacts in the boundary area, and the image artifacts after data augmentation should also be modified. Therefore, we propose a Blend-Net for blending. The Blend-Net takes the pasted image after the paste operation as input to synthesize harmonized final output \( \hat{I}_o \). Since the target is fusion and amending instead of generation, we predict the pixel residual, which is further added back to the input image to composite the final output image. Detailed network structure is illustrated in the supplementary file.

**Synchronization Enhancement.** To further improve the synchronization performance, we introduce a pre-trained sync expert network to boost the model’s performance in synchronization inspired by [26]. The sync expert network consists of two pre-trained encoders, which extract features of image and speech audio within a sliding window, denoted by \( i \) and \( a \) respectively. Different from [26], we only have a short video for training. Therefore, the unsynced speech-image pairs are also exploited to construct the contrastive loss which helps avoid falling into a trivial solution. Its effectiveness is also verified in our experiments. The distance of synced speech-image pairs is desired to be closer while that of the unsynced speech-image pairs should be farther. Thus, we define a contrastive sync loss \( L_s \) based on the metric which is widely used in contrastive learning:

\[
L_s = y \cdot (1 - \cos(i, a)) + (1 - y) \cdot \max(0, \cos(i, a)), \tag{6}
\]

where \( \cos \) is the cosine similarity metric. \( y = 1 \) and \( y = 0 \) represent positive and negative speech-image pairs, respectively. For the input speech, the positive images are the matched images at the same timestamp while the negative images are randomly chosen at some timestamps else.

### 4.5. Training Objectives

Our whole pipeline is trained in a self-supervised manner using the observed frames to provide the supervisory signals. The overall loss includes a canonical mouth image reconstruction loss \( L_m \), a depth loss \( L_d \), a sync loss \( L_s \), and a whole observed image reconstruction loss \( L_w \).

\[
L_m = L_p(\hat{I}_m \circ I^m_{o \rightarrow c}) + \| \hat{I}_c - I^m_{o \rightarrow c} \|_2, \tag{7}
\]

where \( I^m_{o \rightarrow c} \) is the warped ground truth lip image using O2C space mapping (see the supplementary file for more details) and the widely-used perceptual loss \( L_p \) is defined in [40]. Similarly, overall reconstruction loss \( L_w \) is computed as

\[
L_w = L_p(\hat{I}_o, I_o) + \| \hat{I}_o - I_o \|_2. \tag{8}
\]

The canonical head depth map is initialized by the incomplete depth map computed by 3DMM and is trained with the help of the photometric loss as

\[
L_d = \| \hat{I}_o - I_c \|_2, \tag{9}
\]

where \( \hat{I}_o \) is the warped predicted image. With the contrastive sync loss (Sec. 4.4), the overall loss function is

\[
L = \omega_m \cdot L_m + \omega_w \cdot L_w + \omega_d \cdot L_d + \omega_s \cdot L_s. \tag{10}
\]

Implementation details are shown in the supplementary file.

### 5. Experiments

#### 5.1. Datasets and metrics

**Datasets.** We evaluate our algorithm and other methods on three datasets belonging to three specific speakers (Testset I, Testset II and Testset III) as recent speaker-specific
methods [16, 19] usually do. Training videos are collected from [19] as they are publicly available. All video sequences are resampled to 25 FPS and split into training part and test part with a ratio of 90%/10%. The training splits of Testset I (500 × 500) and Testset II (624 × 624) are employed to train the models and then same-identity evaluations are conducted on the test splits. Testset III is utilized for cross-identity, -gender, and -language tests. We present more evaluation results in the supplementary file.

**Evaluation Metrics.** Since both face generation and lip generation models are contained in our quantitative evaluations, we conduct the image quality assessment just around the **mouth region** for a fair comparison. To evaluate the appearance quality, we first utilize image quality metrics Peak Signal-to-Noise Ratio (PSNR) and SSIM [37]. A non-reference objective image sharpness metric based on the Cumulative Probability of Blur Detection (CPBD) [23, 21, 22] is further utilized to measure the overall perceptual image quality. For synchronization evaluation, Landmarks Distance (LMD) around the **mouth region** is exploited to compute the accuracy of lip shape following [6]. The confidence score of lip synchronization computed by a pre-trained SyncNet [12] is also applied to measure the synchronization performance, labeled by Sync.

### 5.2. Comparisons with State of the Arts

**Same-identity Evaluation.** For speaker-independent models, we conduct experiments using their officially-released pre-trained models. Since AD-NeRF [16] and DFRF [28] are speaker-specific models, we retrain them using our training data for a fair comparison. DFRF aims at few-shot learning but its lip synchronization performance witnesses a significant drop when the number of training images decreases from 5000 to 15. Hence, we still train DFRF on thousands of images. Also, we directly use LSP’s [19] pre-trained models on Testset I and Testset II to conduct experiments (subject May for Testset I and Testset III, and subject Obama2 for Testset II). We cannot provide a quantitative comparison with SSP-NeRF [18] here, as neither codes nor pre-trained weights are available. Thus, for comparison, we extract the speech from their released demo and then present qualitative comparisons in the supplementary file.

Quantitative results are shown in Table 1. We divide all the models into two types based on the scale of the training dataset. The upper four algorithms (speaker-independent models) are trained using public video datasets with large amounts of identities and speech-image pairs (e.g., LRS2 [1, 11, 13] and VoxCeleb2 [10]) while the latter four methods (speaker-specific models) only use 5-minute videos of a specific identity for training. Speaker-independent models tend to perform well in synchronization because they can learn lip motions well from various speech-visual pairs in training time. However, the cost is sacrificing the visual quality. The metrics of image quality (e.g., PSNR, SSIM, and CPBD) are much lower than that of specific models. Wav2lip [26] is an exception because it only generates lip images, but the details still can not be produced well (Figure 5). In contrast, speaker-specific models usually have much better image qualities, which is favorable in practice. Our model outperforms all the SOTA algorithms in image quality and achieves the best synchronization performance among all the speaker-specific models, and the Sync score is also competitive when compared to those speaker-independent models. It is also worth mentioning that the Sync scores of Wav2lip [26] and PC-AVS [43] are higher than that of the ground truth. The reason might be that the Sync score is sensitive to synchronization quality only when synchronization quality is within a range and a higher score above a threshold may not necessarily imply a higher synchronization quality. We design a toy experiment to verify it in the supplementary file and our user study can also demonstrate it.

**Cross-identity Evaluation.** Our model also has the abil-
Figure 5. Qualitative results compared with SOTA methods. The Lip area is cropped based on the detected 2D landmarks for a clear comparison. Speaker-independent and speaker-specific models are remarked by different colors.

**Qualitative Results.** Qualitative comparisons of different algorithms are presented in Figure 5. We can see that speaker-independent methods all suffer from low image quality. Besides, ATVG and MakeItTalk have unnatural speech-insensitive movements (e.g., head motion and background motion) and PC-AVS has an extra issue of struggling with identity preservation. Speaker-specific approaches have much better visual quality but most of them can hardly estimate precise lip details when there exist considerable significant lip motions. And fine details like teeth and tongue are not well modeled either. Besides, AD-NeRF sometimes causes head-torso separation due to the adoption of two separate NeRFs. In contrast, our method Speech2Lip performs well in all these situations, as the generated shapes are more accurate and the synthetic images are more clear and realistic, especially the lip area. More results can be found in the supplementary videos.

**Complexity Comparisons.** To further demonstrate the superiority of our model, complexity and computational cost comparisons with speaker-specific models are conducted on Testset I. In Table 2, our model gains the best results with much less complexity and lower computational cost.

### User Study
To verify the perceptual quality, user studies are conducted on 16 generated video clips covering both Testset I and Testset II. Videos are generated by our proposed Speech2Lip and other SOTA algorithms, with algorithm names hidden and video order randomized. 17 participants independently evaluated all the videos. For each video, each participant gave a Mean Opinion Score (MOS) from 1-5 for each of three aspects: qualities of speech-visual synchronization, image fidelity, and image realness. Higher scores represent better quality. The overall results are shown in Table 3. Speech2Lip not only achieves the highest MOS in image quality measurement (e.g., 4.618 and 4.382) but also outperforms all the other algorithms in lip synchronization (e.g., 4.265), exhibiting Speech2Lip has the most satisfactory overall video quality.

### 5.3. Ablation Study
**Contributions of loss functions.** We first conduct an ablation study on loss functions. The evaluation metrics contain the PSNR around the lip area and the sync score. The PSNR of the whole image is also retained since practical applications attach more attention to the overall quality of the talking portrait scene. Results are shown in Table 4, demonstrating the loss functions all contribute to the increased performance on both visual quality and synchronization.
### Table 3. Detailed user study results compared with SOTA methods. The best results are in **bold**.

<table>
<thead>
<tr>
<th>Methods</th>
<th>ATVG</th>
<th>MakeitTalk</th>
<th>Wav2Lip</th>
<th>PC-AVS</th>
<th>LSP</th>
<th>AD-NeRF</th>
<th>DFRF</th>
<th>Speech2Lp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lip Synchronization</td>
<td>1.500</td>
<td>1.853</td>
<td>4.191</td>
<td>3.059</td>
<td>3.118</td>
<td>3.088</td>
<td>2.853</td>
<td><strong>4.265</strong></td>
</tr>
<tr>
<td>Image Fidelity</td>
<td>1.147</td>
<td>2.794</td>
<td>3.265</td>
<td>2.529</td>
<td>3.882</td>
<td>3.235</td>
<td>3.265</td>
<td><strong>4.618</strong></td>
</tr>
<tr>
<td>Image Realness</td>
<td>1.118</td>
<td>2.088</td>
<td>3.765</td>
<td>2.265</td>
<td>3.000</td>
<td>2.324</td>
<td>2.971</td>
<td><strong>4.382</strong></td>
</tr>
</tbody>
</table>

### Table 4. Ablation study results about loss function on Testset I.

<table>
<thead>
<tr>
<th>Methods</th>
<th>PSNR_{lip}↑</th>
<th>PSNR_{img}↑</th>
<th>Sync↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{L}_w )</td>
<td>34.104</td>
<td>36.947</td>
<td>6.313</td>
</tr>
<tr>
<td>( \mathcal{L}_w + \mathcal{L}_d )</td>
<td>34.429</td>
<td>36.994</td>
<td>6.359</td>
</tr>
<tr>
<td>( \mathcal{L}_w + \mathcal{L}_d + \mathcal{L}_m )</td>
<td>34.694</td>
<td>37.193</td>
<td>7.194</td>
</tr>
<tr>
<td>( \mathcal{L}_w + \mathcal{L}_d + \mathcal{L}_m + \mathcal{L}_s )</td>
<td><strong>34.815</strong></td>
<td><strong>37.245</strong></td>
<td><strong>7.771</strong></td>
</tr>
</tbody>
</table>

### Table 5. Ablation study results about network design on Testset I.

<table>
<thead>
<tr>
<th>Methods</th>
<th>PSNR_{lip}↑</th>
<th>PSNR_{img}↑</th>
<th>Sync↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o continuous sampling</td>
<td>34.702</td>
<td>37.214</td>
<td>7.553</td>
</tr>
<tr>
<td>w/o implicit model</td>
<td>34.250</td>
<td>37.058</td>
<td>6.790</td>
</tr>
<tr>
<td>w/o Blend-Net</td>
<td>33.736</td>
<td>36.881</td>
<td>6.065</td>
</tr>
<tr>
<td>w/o time</td>
<td>34.512</td>
<td>37.121</td>
<td>7.146</td>
</tr>
<tr>
<td>Speech2Lp</td>
<td><strong>34.815</strong></td>
<td><strong>37.245</strong></td>
<td><strong>7.771</strong></td>
</tr>
</tbody>
</table>

### Contributions of individual components.
We also explore the benefits of our key components by removing each component to see how the performance changes in Table 5. “w/o implicit model” represents we employ an explicit lip generation model instead, and “w/o time” indicates timestamp is deleted from inputs. From the results, it can be concluded that each design plays an important role. Among them, the Blend-Net significantly improves video quality, and the design of implicit modeling schema is also critical.

### 5.4. Controllable Synthesis Results

Thanks to the joint optimization and completion of a full-head depth map \( D_c \) in GAMEM, our model support synthesizing full-head images driven by novel head poses. Specifically, \( D_c \) can be projected into any observed view to obtain \( D_o \), and \( T_o \) in Eq. (3) are further replaced by \( D_o \) and the new target head pose, respectively. Then, we adopt Eq. (5) to calculate the correspondence. The correspondence is further exploited to warp the synthesized image in the observed space into a new target space. It is noted that depth maps are used for coarse mapping between spaces while facial expression is more taken charge of by fine-grained contents from observed views. The pose-controllable novel view synthesis results are shown in Figure 7. When testing with moderate pose changes, our model achieves comparable results as AD-NeRF.

We note that existing methods with 3D modeling such as AD-NeRF also cannot generate good synthesis results if the evaluated poses deviate too much from the training data. The reason is that the training data widely used in speaker-specific models [19, 16, 28, 18] only show the front view with limited pose variations, being naturally hard for 3D modeling. The 3D mesh is incomplete, and the surface is rough (Figure 6). In addition, the computational costs of 3D methods are extremely high (Table 2).

### 6. Conclusion

In this paper, we propose a novel decomposition-synthesis-composition framework called Speech2Lip for high-fidelity talking head video synthesis, which disentangles speech-sensitive and speech-insensitive motions/appearances. By presenting a synced speech-driven implicit model, a GAMEM module, a Blend-Net, and a contrastive sync loss, we can achieve satisfactory results with only a few minutes of training video. Our framework also supports pose-controllable synthesis. In the future, we plan to study generating realistic expressions driven by speech and explore combining our insights with advanced general image generation methods such as diffusion-based models for better generalizability. We hope that our work inspires more future research in this field and encourages the development of positive applications. However, we also urge caution to prevent any potential abuses. More discussions about limitations are described in the supplementary file.

**Acknowledgement** This work has been supported by the Research Fund from Tencent ARC lab.
References


[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In International Conference on Computer Vision, 2017. 5


[28] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In European conference on computer vision, 2022. 2, 3, 6, 8


