

Perceptually Accurate 3D Talking Head Generation: New Definitions, Speech-Mesh Representation, and Evaluation Metrics

Supplementary Material

Contents

- A Supplementary Video**
- B Emergent Properties of 2D Speech Representation**
- C Speech-Mesh Synchronized Representation**
 - C.1 Network architecture
 - C.2 Training pipeline
 - C.3 Dataset statistics
- D Details of Human Study on Lip Synchronization Criteria**
- E Evaluation Metrics**
 - E.1 Definition and implementation details
 - E.2 Human study on perceptual metric
- F Implementation Details of Ablation Study**
- G Additional Results**
 - G.1 Human study on applying perceptual loss
 - G.2 FDD evaluation on applying perceptual loss
 - G.3 Qualitative result of temporal synchronization
 - G.4 Stability comparison on loss and cosine similarity
- H Discussion**

A. Supplementary Video

This work focuses on 3D facial motions, which are best viewed in video format. Please refer to the attached **supplementary video**. The video contains qualitative results of lip synchronization on the VOCASET and MEAD-3D test sets, demonstrating the effectiveness of our method in enhancing lip synchronization in aspects of lip readability and expressiveness.

B. Emergent Properties of 2D Speech Representation

In this section, we conduct further analyses of 2D speech representation (*i.e.*, 2D prior knowledge), which motivate the transfer of the emergent properties of 2D speech representation to the speech-mesh representation space using a curriculum learning approach.

We observe that the 2D audio-visual speech representation, trained with a transformer architecture and an extensive video dataset [1], inherently exhibits the desirable properties for lip synchronization that we aim to achieve. We visualize a cosine similarity versus temporal offset graph and a t-SNE visualization of the 2D audio-visual speech representation in Fig. S1. The speech representation exhibits the properties regarding the critical aspects of lip synchronization:

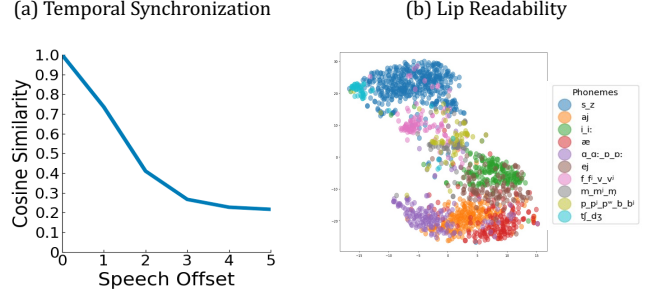


Figure S1. Emergent properties of 2D speech representation. We visualize a cosine similarity versus temporal offset graph and a t-SNE visualization of the 2D audio-visual speech representation. The 2D speech representation already possesses desired properties we pursue, which motivates us to transfer the emergent properties to the speech-mesh representation space.

(1) Temporal sensitivity in Fig. S1-(a), (2) clear separation and clustering of speech features corresponding to the same phoneme group in Fig. S1-(b), and (3) a directional progression of speech features as intensity increases from the lowest to the highest levels in Fig. 5-(b) of the main paper¹. This motivates us to transfer these emergent properties to the 3D speech-mesh representation through the curriculum learning approach, as mentioned in Sec. 4 of the main paper. Furthermore, as shown in Figs. 4 and 5 of the main paper, we demonstrate that these properties are successfully transferred to the speech-mesh representation.

C. Speech-Mesh Synchronized Representation

We provide more details on the network architecture of audio-visual speech representation and speech-mesh representation (Sec. C.1). In addition, we provide the training details of the two-stage training process (Sec. C.2) and dataset statistics of speech-mesh benchmark datasets (Sec. C.3).

C.1. Network architecture

To improve the reproducibility of our speech-mesh representation, we further illustrate the detailed network architectures for the audio-visual speech representation and the speech-mesh representation, which are shown in Table S1.

¹We freeze the pre-trained speech encoder from stage 1 and utilize it as the speech encoder in stage 2, which ensures that the speech representation in both stages shares the same favorable property of expressiveness.

Stage	Module	Input \rightarrow Output	Layer Operation
1	Speech Tokenizer	$\mathbf{X}_s(C_s, H_s, W_s) \rightarrow \mathbf{S}(N, H)$	Conv2D((1, 16), (1, 16), H)
	Speech Encoder	$\mathbf{S}^{unmask}(N^{unmask}, H) \rightarrow \mathbf{Z}_s(N^{unmask}, H)$	$[\text{MHSA}(H, 8) \rightarrow \text{FFN}(H)] \times 10 \rightarrow \text{LN}$
	Speech Decoder	$\mathbf{F}'_s \rightarrow \hat{\mathbf{S}}(N^{mask}, C_s \cdot H_s \cdot W_s)$	Concat(Linear($H, 384$) + PE(N^{unmask}), PE(N^{mask})) \rightarrow MHSA($384, 8$) \rightarrow FFN($384 \cdot 4$) \rightarrow [MHSA($384, 8$) \rightarrow MHCA($Z_s, 384, 6$) \rightarrow FFN($384 \cdot 4$)] $\times 3 \rightarrow$ LN \rightarrow Linear($C_s \cdot H_s \cdot W_s$) \rightarrow Slice[$N^{unmask} :$]
	Video Tokenizer	$\mathbf{X}_v(C_v, T, H_v, W_v) \rightarrow \mathbf{V}(M, H)$	Conv3D((1, 16, 16), (1, 16, 16), H)
	Video Encoder	$\mathbf{V}^{unmask}(M^{unmask}, H) \rightarrow \mathbf{Z}_v(M^{unmask}, H)$	$[\text{MHSA}(H, 8) \rightarrow \text{FFN}(H)] \times 10 \rightarrow \text{LN}$
	Video Decoder	$\mathbf{F}'_v \rightarrow \hat{\mathbf{V}}(M^{mask}, C_v \cdot H_v \cdot W_v)$	Concat(Linear($H, 384$) + PE(M^{unmask}), PE(M^{mask})) \rightarrow MHSA($H, 8$) \rightarrow FFN(H) \rightarrow [MHSA($H, 8$) \rightarrow MHCA($Z_v, H, 6$) \rightarrow FFN(H)] $\times 3 \rightarrow$ LN \rightarrow Linear($C_v \cdot H_v \cdot W_v$) \rightarrow Slice[$M^{unmask} :$]
	Fusion Encoder	$\mathbf{Z}_s, \mathbf{Z}_v \rightarrow \mathbf{F}_s(N^{unmask}, H)$ $\mathbf{Z}_s, \mathbf{Z}_v \rightarrow \mathbf{F}_v(M^{unmask}, H)$	$[\text{MHSA}(H, 8) \rightarrow \text{MHCA}(Z_v, H, 8) \rightarrow \text{FFN}(H \cdot 4)] \times 2$ $[\text{MHSA}(H, 8) \rightarrow \text{MHCA}(Z_s, H, 8) \rightarrow \text{FFN}(H \cdot 4)] \times 2$
2	Mesh Tokenizer	$\mathbf{X}_m(T, V \cdot 3) \rightarrow \mathbf{M}(T, H)$	Linear(H)
	Mesh Encoder	$\mathbf{M} \rightarrow \mathbf{Z}_m(T, H)$	$[\text{MHSA}(H, 8) \rightarrow \text{FFN}(H)] \times 10 \rightarrow \text{LN}$

Table S1. Architecture details. The parameters of network architectures. Conv2D(k, s, n) denotes a 2D Convolutional layer with kernel size k , stride size s , and output channel of n . MHSA($d, nhead$) denotes a multi-head self-attention layer with the input channels d and the number of heads in multi-head attention $nhead$. MHCA($ca, d, nhead$) denotes a multi-head cross-attention layer with additional cross-attention input ca . PE(a) is a position embedding layer where a denotes the length of the position vector. FFN(d) is a feed-forward layer. Linear(n) denotes a linear layer with output channels of n . LN denotes layer normalization and Slice[$s :$] denotes slice operation.

C.2. Training pipeline

Two-stage training process. In our experiment, we set $T = 5$, $H = 512$, and $P = 30$. For training the audio-visual speech representation, we use $C_s = 1$, $H_s = 64$, $W_s = 128$, $N = 512$ for speech modality and $C_v = 3$, $H_v = 160$, $W_v = 160$, $M = 500$ for video modality. We train the audio-visual speech representation using LRS on two NVIDIA A6000 for 100 epochs with the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.95$ and $\epsilon = 1e-8$), where the learning rate is initialized as $3e-4$, and the mini-batch size is set as 40. For training the speech-mesh representation, we use the number of vertices $V = 5023$. We train the speech-mesh representation using LRS-3D with the mini-batch size of 80, and other hyper-parameters remain unchanged as Stage 1.

Perceptual loss. We employ our speech-mesh representation as a perceptual loss to enhance the perceptual accuracy of the 3D talking head model. We finetune our speech-mesh representation using the VOCASET [3] train split on an NVIDIA A6000 for 5 epochs with the initial learning rate $1e-4$ and other hyper-parameters remain unchanged as Stage 2. To train the 3D talking head models with our perceptual loss, we split the generated mesh from the model into 5 frames using a sliding window size of 1. We make a batch of size 80 and get uni-modal embeddings from our representation. We

Dataset	# Vertex clips	# Speaker IDs	Total hours	FPS
VOCASET	475	12	0.5	30
BIWI	1109	14	1.4	25
LRS3-3D	17752	788	61.1	25
MEAD-3D	8765	15	10.2	30

Table S2. Statistics of speech-mesh paired benchmark. We use VOCASET, LRS3-3D and MEAD-3D speech-mesh paired datasets in our experiments. We construct two large-scale speech-mesh benchmark datasets, LRS3-3D and MEAD-3D, using monocular face reconstruction methods.

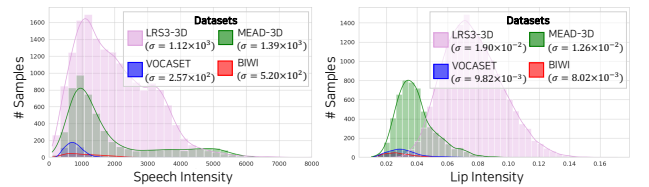


Figure S2. Speech and lip intensity distributions across datasets. We present speech and lip intensity distributions and corresponding standard deviation values across datasets.

additionally apply the InfoNCE loss with a weight of $1e-7$ to the original training loss of the model.

C.3. Dataset statistics

We construct LRS3-3D and MEAD-3D by processing LRS3 [1] and MEAD [12] videos using two monocular face reconstruction methods, respectively: SPECTRE [6] for LRS3, which ensures accurate lip movements, and SMIRK [9] for MEAD, which captures diverse speech and lip movement intensities. We construct a test split for LRS-3D, involving 934 clips. We split MEAD-3D to construct a test split, which includes 3470 clips.

Table S2 and Fig. S2 show the statistics of the existing (VOCASET [3], BIWI [5]) and the newly proposed large-scale speech-mesh benchmark datasets (LRS3-3D and MEAD-3D). As shown in Table S2, LRS3-3D and MEAD-3D have notably larger data sizes than VOCASET and BIWI. Fig. S2 presents the broader speech and lip intensity² distributions of LRS3-3D and MEAD-3D with higher standard deviations (σ), indicating greater variability in facial motions. In contrast, VOCASET and BIWI show limitations in both scale and diversity.

D. Details for Human Study on Lip Synchronization Criteria

Human preference between the speech and lip intensities. We conduct a preliminary experiment to demonstrate the positive correlation of human preference between the intensity of speech and lip movements in the 3D talking face field. Using the intensity annotations from the MEAD dataset [12], we first split the MEAD-3D dataset into three categories: Level 1, Level 2, and Level 3, representing different intensity levels. Then, we train a 3D talking face model [4] using VOCASET [3] (to ensure the quality of generation) and each intensity split separately. This results in three distinct models, each of which tends to generate lip movements biased toward the intensity level present in its training data, regardless of the speech intensity provided as input. We input three speeches with intensity levels ranging from Level 1 to Level 3 into each of the three biased models, producing nine intensity configurations in the generated mesh sequences as shown in Tab.1-[Left] of the main paper. We then asked 17 participants, a balanced group of males and females from a non-expert background in the field, to rank their preferences in three videos, assigning a score from 1 (least preferred) to 3 (most preferred). Each video has the same speech (identical in utterance and intensity) but differs in the intensity of the lip movements.

Human preference on Temporal sync. vs. Expressiveness. We design a simple A/B test to investigate an interesting aspect of human perception for lip synchronization. We use the two biased models from the previous human study: one

trained to generate Level 1 lip movements and the other trained to generate Level 3 lip movements, regardless of the speech intensity. For each model, we create two types of samples. Sample A is temporally synchronized but lacks expressive synchronization (*e.g.*, speech of Level 3 intensity and lip movements of Level 1 intensity). In contrast, sample B has expressive synchronization (*e.g.*, speech of Level 3 intensity and lip movements of Level 3 intensity) but is temporally misaligned. To introduce the temporal mismatch in Sample B, we make the speech lead the lip movements by 100ms, which exceeds twice the established maximum acceptable synchrony [11]. We then asked 28 participants, comprising a balanced group of males and females from a non-expert background in the field, to choose which sample they prefer based on how well the lip movements correspond to the speech in sample A vs. B.

E. Evaluation Metrics

We present the comprehensive definitions of the evaluation metrics and their implementation details (Sec. E.1). In addition, we provide the human study on the perceptual metric (Sec. E.2), which demonstrates the correlation between our perceptual metric and human preference.

E.1. Definition and implementation details

Mean Temporal Misalignment (MTM). Let $\mathbf{V}(t)$ represent the ground truth vertex sequences, where each frame t consists of vertex positions $\mathbf{v}_t \in \mathbb{R}^{N \times 3}$, with N being the number of vertices. Similarly, $\hat{\mathbf{V}}(t)$ represents the predicted vertex sequences, with predicted vertex positions $\hat{\mathbf{v}}_t \in \mathbb{R}^{N \times 3}$. For each sample k , we select two specific vertices that correspond to the center of the upper and lower lips, extracting the upper-lip vertex sequence $\mathbf{V}_u(t) \in \mathbb{R}^{T \times 3}$ and the lower-lip vertex sequence $\mathbf{V}_l(t) \in \mathbb{R}^{T \times 3}$ (refer to Fig. S3).

We then calculate the Euclidean distance between the upper and lower lip vertices over time to derive the ground truth lip distance sequence $d_v(t) = \|\mathbf{V}_u(t) - \mathbf{V}_l(t)\|$. The same process is applied to obtain the predicted lip distance sequence $\hat{d}_v(t)$. To reduce noise, we apply a Gaussian filter to both lip distance sequences.

Next, we compute the first-order derivatives of the smoothed lip distance sequences to capture the dynamic changes in lip movement. We then use Derivative Dynamic Time Warping (DDTW) [7] to determine the optimal alignment path $\mathcal{A} = \{(i, j)\}$ between the derivative sequences $\delta \tilde{d}_v(t)$ and $\delta \hat{\tilde{d}}_v(t)$. We identify local extrema (peaks and valleys) in each derivative sequence and match only extrema of the same type (*i.e.*, both maxima or both minima) to compute the absolute time difference $\delta t_n = |i - j|$ (refer to Fig. S4).

For each sample k , the sample mean temporal misalignment Δt_k is computed as $\Delta t_k = \frac{1}{M} \sum_{m=1}^M \delta t_n$, where M is the number of matched extrema pairs in the sample.

²Lip intensity was normalized by eye distance to account for differences between FLAME and BIWI topologies.

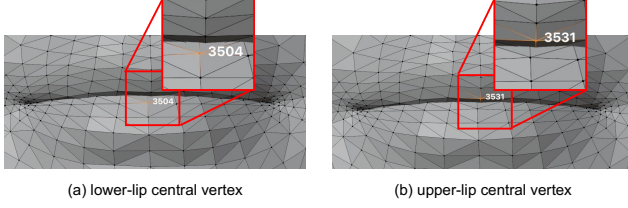


Figure S3. Central vertices of the lower and upper lips. We select two specific vertices that correspond to the center of the upper and lower lips to extract the lip vertex displacement sequences.

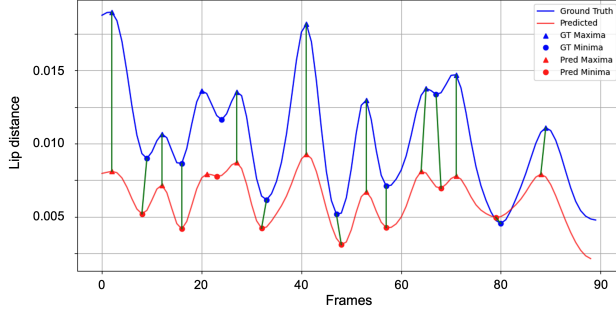


Figure S4. An example of DDTW matching results between ground truth and predicted lip distance sequences. We present an example of the DDTW local extrema correspondences of the ground truth and predicted lip vertex displacement sequences. We represent matched local extrema using green lines.

Finally, the overall mean temporal misalignment is given by $\overline{\Delta t} = \frac{1}{K} \sum_{k=1}^K \Delta t_k$, where K is the total number of samples. A smaller $\overline{\Delta t}$ indicates better temporal alignment of the predicted sequences with the ground truth lip movements. To express the Mean Temporal Misalignment (MTM) in milliseconds, we multiply $\overline{\Delta t}$ by the frame duration for the given dataset. For instance, for a dataset with 25 FPS, the MTM is obtained by multiplying $\overline{\Delta t}$ by 40ms. Refer to Algorithm 1 for more details on the MTM calculation. Furthermore, to validate the physical accuracy of our proposed temporal synchronization metric, we present a graph showing the relationship between the temporal offset and the corresponding MTM values. Specifically, we introduce temporal mismatch to the ground truth mesh sequences of VOCASET [3] by making the speech leading the mesh sequences by 0 to 10 frames (*i.e.*, 0 to 333ms for VOCASET). Figure S5 shows that MTM accurately captures the degree of temporal mismatch across the samples, demonstrating the effectiveness and physical accuracy of our proposed temporal synchronization metric.

Perceptual Lip Readability Score (PLRS). We train speech-mesh representation using our proposed two-stage training process with different datasets, initializations, and batch sizes. For both Stage 1 and Stage 2, we use a batch size

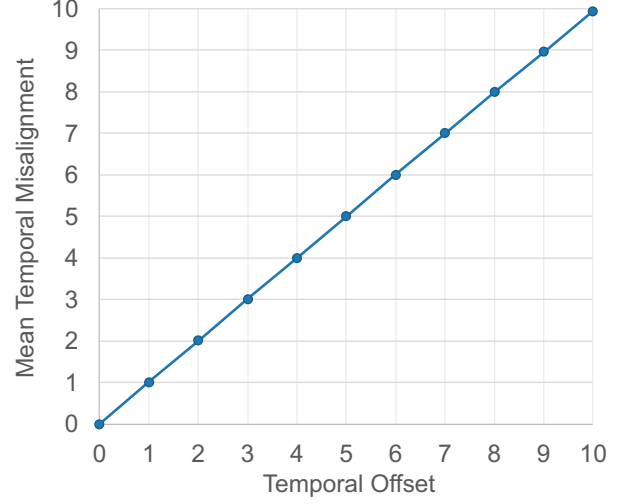


Figure S5. Physical accuracy of Mean Temporal Misalignment. We introduce temporal mismatch to the ground truth mesh sequences of VOCASET [3] by shifting the speech to lead the mesh sequences by 0 to 10 frames (where 0 represents no mismatch). For each temporal offset, we calculate the average MTM and plot a graph showing the relationship between the temporal offset and the corresponding MTM values.

of 256. Given a speech and generated mesh pair $(\mathbf{X}_s, \hat{\mathbf{X}}_m)$, we split the generated mesh into 5 frames with a sliding window size of 5 to make mesh tokens $\{\hat{\mathbf{M}}_i\}_{i=1}^G$, and the speech is also converted into corresponding speech tokens $\{\mathbf{S}_i\}_{i=1}^G$. We then compute the average cosine similarity between mean pooled speech embeddings $\{\mathbf{c}_{s,i}\}_{i=1}^G$ and mesh embeddings $\{\mathbf{c}_{m,i}\}_{i=1}^G$:

$$PLRS(\mathbf{S}, \hat{\mathbf{M}}) = \frac{1}{G} \sum_{i=1}^G \frac{\mathbf{c}_{s,i} \cdot \mathbf{c}_{m,i}}{\|\mathbf{c}_{s,i}\| \|\mathbf{c}_{m,i}\|}. \quad (a)$$

Speech-Lip Intensity Correlation Coefficient (SLCC). First, we define speech intensity using speech loudness, specifically the Root Mean Square (RMS) value, which is a widely accepted measure of speech intensity in signal processing. RMS loudness effectively captures the energy of the speech signal and provides an accurate representation of perceived speech intensity. However, since RMS values can vary based on recording conditions (*e.g.*, microphone gain and distance from the microphone), we perform identity-wise z-normalization on the RMS values to standardize them, assuming that clips belonging to the same identity are recorded under similar conditions. The Speech Intensity (SI) is thus defined as:

$$SI_k = \frac{RMS_k - \mu_{s,i}}{\sigma_{s,i}}, \quad (b)$$

where RMS_k is the averaged RMS value of k-th video clip and $\mu_{s,i}$ and $\sigma_{s,i}$ are the mean and standard deviation of the

speech RMS values for the clips with identity $i \in I$.

To define Lip Intensity (LI), we first measure the averaged lip displacement value of k -th video clip Dist_k . as:

$$\text{Dist}_k = \sqrt{\frac{1}{T_k - 1} \sum_{t=1}^{T_k-1} \left(\frac{1}{V_l} \sum_{v=1}^{V_l} \|\mathbf{l}_{t+1,v} - \mathbf{l}_{t,v}\| \right)^2}, \quad (\text{c})$$

where T_k is the number of frames in clip k , V_l is the number of vertices in the lip region, and $\mathbf{l}_{t,v} \in \mathbb{R}^3$ represents a vertex position in the lip region at time t . Similar to Speech Intensity, we perform identity-wise z-normalization to the lip displacement values to mitigate individual bias in lip movement as:

$$\text{LI}_k = \frac{\text{Dist}_k - \mu_{l,i}}{\sigma_{l,i}}, \quad (\text{d})$$

where $\mu_{l,i}$ and $\sigma_{l,i}$ are the mean and standard deviation of the lip displacement values for the clips belonging to identity $i \in I$.

Finally, we can obtain the Speech and Lip Correlation Coefficient as:

$$r_{SL} = \frac{\sum_{k=1}^K (SI_k - \bar{SI})(LI_k - \bar{LI})}{\sqrt{\sum_{k=1}^K (SI_k - \bar{SI})^2} \sqrt{\sum_{k=1}^K (LI_k - \bar{LI})^2}}, \quad (\text{e})$$

where $\bar{SI} = \frac{1}{K} \sum_{k=1}^K SI_k$ and $\bar{LI} = \frac{1}{K} \sum_{k=1}^K LI_k$.

E.2. Human study on perceptual metric

To validate that our proposed perceptual metric, Perceptual Lip Readability Score (PLRS), effectively evaluates perceptual alignment, we conduct a human study that assesses the correlation between the metric scores and human preferences. We collect meshes from the ground-truth VOCASET [3] dataset and those generated by FaceFormer [4], CodeTalker [13] and SelfTalk [8]. We measure the PLRS and the existing evaluation metric Lip Vertex Error (LVE) for the generated meshes of each model, and subsequently rank the models by their PLRSs and LVEs. We ask 16 participants, evenly balanced in gender and from non-expert backgrounds, to rank the models based on their preferences. We then compute the Spearman’s correlation coefficient ρ to compare the PLRS rankings and the LVE rankings with the human preference rankings. As shown in Table S3, PLRS exhibits a far more positive correlation with human preferences compared to the LVE. This highlights the efficacy of our proposed metric in evaluating perceptual lip readability from a human perspective.

F. Implementation Details of Ablation Study

In this section, we provide implementation details of model variants ablated from our speech-mesh representation: the 3D SyncNet and the representation w/o 2D prior.

Metric	Spearman’s ρ
LVE	0.166
PLRS	0.437

Table S3. Human study on perceptual metric. We conduct a human study to validate our proposed perceptual metric, PLRS. We compute the Spearman’s correlation coefficient ρ to compare the PLRS rankings with the human preference rankings.

3D SyncNet. Inspired by Chung *et al.* [2], we train 3D SyncNet to evaluate the performance of our transformer-based model compared to a CNN-based model. 3D SyncNet is trained using InfoNCE loss with a batch size of 80. The architecture of 3D SyncNet consists of the mesh encoder comprising three dilated convolutional layers and the speech encoder with six convolution layers followed by two linear layers. The mesh and speech features are extracted from each encoder, respectively. We train 3D SyncNet on an NVIDIA RTX 3090 GPU for 20 epochs using LRS3-3D. Also, for imposing the perceptual loss to 3D talking head models with 3D SyncNet, we finetune the model with VOCASET [3] train split for 5 epochs, as our speech-mesh representation model does.

Ours w/o 2D prior. We train speech-mesh representation without Stage 1 training to evaluate the effectiveness of our learned 2D prior. We train the speech encoder and mesh encoder, both with the same architecture as Stage 2, and the other hyperparameters are the same as in Stage 2.

G. Additional Results

In this section, we present quantitative results on human studies (refer to Sec. G.1) and Upper Face Dynamics Deviation (FDD) evaluation (refer to Sec. G.2), comparing samples generated by the base models [4, 8, 13] with and without perceptual loss to demonstrate the effectiveness of our speech-mesh representation. Additionally, we provide the qualitative result of temporal synchronization for the base models [4, 13] (refer to Sec. G.3). We also provide comparisons on the stability of perceptual loss and cosine similarity for ablated model variants (refer to Sec. G.4).

G.1. Human study on applying perceptual loss

We conduct a human study to evaluate the perceptual preference for our method with two configurations: (1) training and testing on VOCASET, and (2) training on the combined MEAD-3D and VOCASET and testing on MEAD-3D, as mentioned in Sec. 6.1 of the main paper.

In the first configuration, we ask participants, evenly balanced group of males and females with non-expert backgrounds, to compare two videos: one generated by the base model [4, 8, 13] without our perceptual loss and the other with it. To assess the quality of generated meshes, we design

two separate descriptions—one focusing on lip synchronization and the other on overall quality. For lip synchronization, participants are provided with the following description: “Please evaluate the lip synchronization between the speech and the lip movements in videos A and B, and choose the one that is more realistic and preferred.” A total of 18 participants take part in this evaluation. Table S4 shows that the participants significantly favor the models incorporating our perceptual loss with an overall preference rate of 72.9%. For overall quality, the description is as follows: “Please evaluate the overall quality of facial movements in videos A and B, and choose the one that is more realistic and preferred.” This evaluation involves 15 participants. As shown in Table S5, the participants show a strong preference for the model incorporating perceptual loss, with an overall preference rate of 73.3%, indicating that the perceptual loss not only improves lip synchronization but also enhances the overall quality of facial movements.

In the second configuration, we ask 14 participants, also an evenly balanced group of males and females with non-expert backgrounds, to compare three videos: one generated by the base model [4, 8, 13] trained on VOCASET, another generated by the base model trained on both MEAD-3D and VOCASET without our perceptual loss, and the other generated by the base model trained on both MEAD-3D and VOCASET with our perceptual loss. The description is as follows: “Please rate the lip synchronization between the speech and the lip movements in videos A through C, with 3 being the most realistic and preferred, and 1 being the least.” As indicated in Table S6-(a) and (b), the participants significantly prefer the models incorporating MEAD-3D and our perceptual loss each by in 76.9% and 67.9% overall. Notably, incorporating both MEAD-3D dataset and the perceptual loss results in 84.6% of participants favoring the model, as shown in Table S6-(c), compared to the original models.

This preference on the two configurations highlights the effectiveness of our speech-mesh representation as a plug-in module in enhancing lip synchronization from the perspective of human perception.

G.2. FDD evaluation on applying perceptual loss

In Table S7, we measure Upper Face Dynamics Deviation (FDD) [13], a widely used metric for the upper face evaluation, to assess the effectiveness of our perceptual loss. The models applying our perceptual loss achieve similar or improved FDD scores. It is expected because FDD is not the main focus of our work due to no direct relationship with the quality of lip movements.

G.3. Qualitative result of temporal synchronization

We present the qualitative result of temporal synchronization using existing base models [4, 8, 13] (See Fig. S8). Given

Model	w/o Our rep.	w/ Our rep.
FaceFormer	13.7%	86.3%
CodeTalker	32.4%	67.6%
SelfTalk	35.3%	64.7%
Overall	27.1%	72.9%

Table S4. Human study results on lip synchronization in configuration 1. We adopt A/B test and report the percentage (%) of preferences for A (Ours) over B, assessing the generated meshes on lip sync. Participants significantly favor the models incorporating our perceptual loss by in overall 72.9%.

Model	w/o Our rep.	w/ Our rep.
FaceFormer	14.4%	85.6%
CodeTalker	27.8%	72.2%
SelfTalk	37.8%	62.2%
Overall	26.7%	73.3%

Table S5. Human study results on overall quality in configuration 1. We adopt A/B test and report the percentage (%) of preferences for A (Ours) over B, assessing the generated meshes on overall quality. Participants show a strong preference for the models applying our perceptual loss, with an overall preference rate of 73.3%.

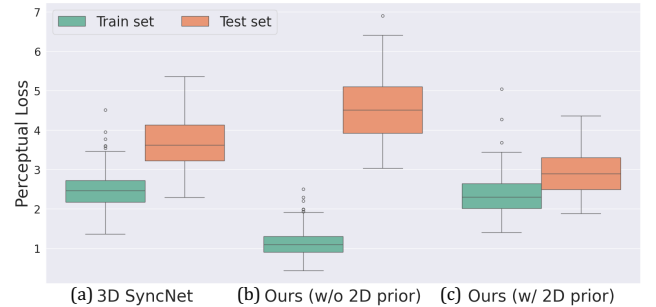


Figure S6. Perceptual loss stability. We visualize the perceptual loss between GT speech-mesh pairs on VOCASET samples. Our representation demonstrates strong generalization capability and provides a stable training signal compared to 3D SyncNet and our representation without 2D prior.

rendered 3D face mesh sequences, we place a vertical line with two pixel points near the lip region and extract the y-t slices of the mesh sequences to visualize the timing of lip closure and opening. Next, we align the y-t slices with their corresponding speech waveforms and mel-spectrograms along the time axis. We observe that these models already have a reasonable temporal synchronization capability. Specifically, the timing of lip closure (*e.g.*, for the /p/ sound) in the y-t slices aligns with minimal amplitude in both the speech waveforms and mel-spectrogram, while the timing of lip

Model	(a)		(b)		(c)	
	Original	Original + MEAD-3D	Original + MEAD-3D	Original + MEAD-3D + Our rep.	Original	Original + MEAD-3D + Our rep.
FaceFormer	33.3%	66.7%	32.1%	67.9%	19.2%	80.8%
CodeTalker	17.9%	82.1%	34.6%	65.4%	19.0%	91.0%
SelfTalk	17.9%	82.1%	29.5%	70.5%	17.9%	82.1%
Overall	23.1%	76.9%	32.1%	67.9%	15.4%	84.6%

Table S6. Human study results on lip synchronization in configuration 2. We report the percentage (%) of preferences for A over B, assessing the generated meshes on lip sync. Overall 84.6% of participants prefer the model with MEAD-3D and our perceptual loss.

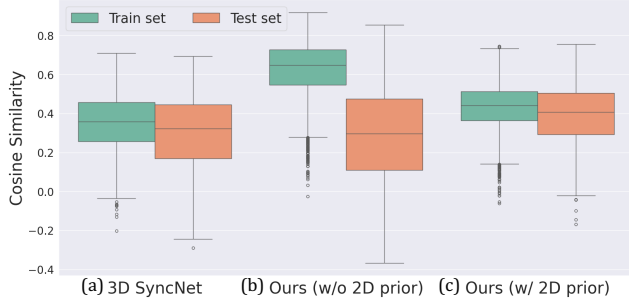


Figure S7. Cosine similarity stability. We visualize the cosine similarity between GT speech-mesh pairs on VOCASET samples. Our representation demonstrates strong generalization capability compared to 3D SyncNet and our representation without 2D prior.

	FDD ↓ ($\times 10^{-7}$ mm)
FaceFormer	3.789
+ Ours rep.	3.325
CodeTalker	3.414
+ Ours rep.	3.259
SelfTalk	3.319
+ Ours rep.	3.424

Table S7. FDD evaluation. We report Upper Face Dynamics Deviation (FDD) scores to evaluate the variation in upper facial dynamics, which is not the main focus of our work. As expected, the models trained with our perceptual loss show similar or improved FDD scores.

opening (*e.g.*, for the /r/ sound) in the y-t slices coincides with a large amplitude in both speech representations.

G.4. Stability comparison on loss and cosine similarity

To utilize our speech-mesh synchronized representation as a perceptual loss, it is essential to provide a stable training signal to the 3D talking head model. In the domain of 2D audio-visual speech representation, Yaman *et al.* [14] reveal that the transformer-based architecture [10] learns more robust representation and provides more stable guidance to talking head models compared to a CNN-based approach [2].

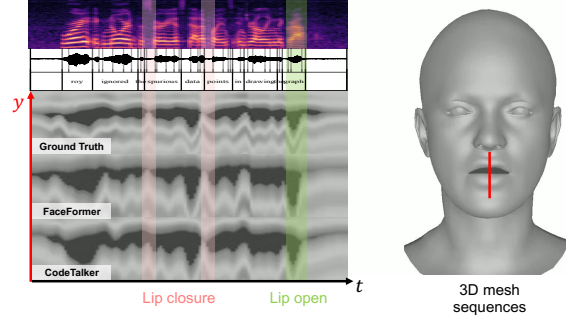


Figure S8. Qualitative results of temporal synchronization on existing models. We plot y-t slices of rendered 3D face mesh sequences on the lip region with corresponding speech waveforms and mel-spectrogram. We also indicate the time steps of lip closure and opening with vertical lines. This implies that existing models already exhibit reasonable temporal sync. capability.

To explore whether these observations hold for 3D speech-mesh representations, we evaluate both the lip-sync loss and cosine similarity across 3D SyncNet, our representation without 2D prior and our final representation. This analysis aims to validate the effectiveness of the transformer-based architecture and curriculum learning with a pre-trained 2D speech representation.

Specifically, we measure the perceptual loss and cosine similarity, computing the mean and standard deviation for both the train and test samples. Figures S6 and S7 show the comparisons of perceptual loss and cosine similarity comparison across the three representation variants. We denote the train samples as green box plots and test samples as orange box plots, respectively.

Our speech-mesh representation (Figs. S6-(c) and S7-(c)) demonstrates the highest stability, exhibiting the lowest standard deviations (the height of the box plots) on test set in both lip-sync loss and cosine similarity. In contrast, the representation without 2D prior (Figs. S6-(b) and S7-(b)) reveals significant discrepancies between the train and test samples on both the lip-sync loss and cosine similarity, indicating poor generalization capability. Additionally, it shows the highest standard deviations, which potentially cause unstable training. Meanwhile, 3D SyncNet (Figs. S6-(a) and S7-(a)) displays the worst mean values of perceptual

	Time ↓ (sec.)	Mem. ↓ (MB)
FaceFormer	0.447	1461
+ Ours rep.	0.537	1738
CodeTalker	0.138	3393
+ Ours rep.	0.289	3675
SelfTalk	0.175	8204
+ Ours rep.	0.320	8480

Table S8. Training efficiency. We compared the memory consumption and single-iteration speed during training with and without the perceptual loss.

loss and cosine similarity among the three.

H. Discussion

Limitations. While our perceptual loss is applied only during training, which ensures that the resource requirements at inference remain unchanged, it requires additional computational resources during training. In Table S8, we compare memory consumption and single-iteration speed during training, measured on a single A6000 GPU. Also, to capture the intricate correspondence between speech and 3D face mesh, we construct large-scale speech-mesh paired datasets, LRS3-3D and MEAD-3D. To this end, we utilize state-of-the-art monocular face reconstruction methods [6, 9], which may impose limitations on the quality of the 3D mesh in the reconstructed datasets.

Ethical considerations. Our method can generate realistic 3D talking faces from arbitrary audio signals, relying on both the 3D scan data collected from actors and the reconstructed data from 2D talking videos. Thus, while this technology has powerful applications, it also poses risks of misuse, such as creating harmful or embarrassing content. To mitigate these risks, we emphasize raising public awareness and promoting ethical and responsible use through continued research.

Algorithm 1 Mean Temporal Misalignment Calculation

Require: GT vertex sequence $V(t)$, Predicted vertex sequence $\hat{V}(t)$
Ensure: Overall mean temporal misalignment $\overline{\Delta t}$

```

1: Initialize list of sample mean misalignments:  $\{\Delta t_k\} \leftarrow \emptyset$ 
2: for each sample  $k$  do
3:   Initialize time differences list:  $\{\delta t_n\} \leftarrow \emptyset$ 
4:   Extract lip vertices:
5:     Upper lip vertex  $V_u(t) \in \mathbb{R}^3$  from  $V(t)$ 
6:     Lower lip vertex  $V_l(t) \in \mathbb{R}^3$  from  $V(t)$ 
7:     Predicted upper lip vertex  $\hat{V}_u(t) \in \mathbb{R}^3$  from  $\hat{V}(t)$ 
8:     Predicted lower lip vertex  $\hat{V}_l(t) \in \mathbb{R}^3$  from  $\hat{V}(t)$ 
9:   Compute lip distance sequences:
10:     $d_v(t) = \|V_u(t) - V_l(t)\|$ 
11:     $\hat{d}_v(t) = \|\hat{V}_u(t) - \hat{V}_l(t)\|$ 
12:   Smooth sequences using Gaussian filter:
13:     $\tilde{d}_v(t) = \text{Gauss}(d_v(t))$ 
14:     $\tilde{\hat{d}}_v(t) = \text{Gauss}(\hat{d}_v(t))$ 
15:   Compute derivatives:
16:     $\delta \tilde{d}_v(t) = \tilde{d}_v(t) - \tilde{d}_v(t-1)$ 
17:     $\delta \tilde{\hat{d}}_v(t) = \tilde{\hat{d}}_v(t) - \tilde{\hat{d}}_v(t-1)$ 
18:   Perform DDTW to find alignment path  $\mathcal{A} = \{(i, j)\}$ 
19:   Identify local extrema in  $\tilde{d}_v(t)$  and  $\tilde{\hat{d}}_v(t)$ 
20:   for each aligned pair  $(i, j) \in \mathcal{A}$  do
21:     if  $i$  and  $j$  are matching extrema of same type then
22:       if  $j$  is within neighboring extrema range of  $i$  in  $\tilde{d}_v(t)$  then
23:         Compute time difference:  $\delta t_n \leftarrow |i - j|$ 
24:         Append  $\delta t_n$  to  $\{\delta t_n\}$ 
25:       end if
26:     end if
27:   end for
28:   if  $\{\delta t_n\} \neq \emptyset$  then
29:     Compute mean delta time for clip  $k$ :
30:      $\Delta t_k = \frac{1}{N} \sum_{n=1}^N \delta t_n$ 
31:     Append  $\Delta t_k$  to  $\{\Delta t_k\}$ 
32:   end if
33: end for
34: if  $\{\Delta t_k\} \neq \emptyset$  then
35:   Compute overall mean temporal misalignment:
36:    $\overline{\Delta t} = \frac{1}{K} \sum_{k=1}^K \Delta t_k$ 
37: else
38:    $\overline{\Delta t}$  is undefined
39: end if

```


References

- [1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. [1](#), [3](#)
- [2] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. [5](#), [7](#)
- [3] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *CVPR*, 2019. [2](#), [3](#), [4](#), [5](#)
- [4] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *CVPR*, 2022. [3](#), [5](#), [6](#)
- [5] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE TMM*, 2010. [3](#)
- [6] Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Spectre: Visual speech-informed perceptual 3d facial expression reconstruction from videos. In *CVPRW*, pages 5745–5755, 2023. [3](#), [8](#)
- [7] Eamonn J. Keogh and M. Pazzani. Derivative dynamic time warping. In *In First SIAM International Conference on Data Mining*, 2001. [3](#)
- [8] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *ACM MM*, 2023. [5](#), [6](#)
- [9] George Retsinas, Panagiotis P Filntisis, Radek Danecek, Victoria F Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *CVPR*, 2024. [3](#), [8](#)
- [10] Bowen Shi, Wei-Ning Hsu, Kushal Lakhota, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *ICLR*, 2022. [7](#)
- [11] Argiro Vatakis and Charles Spence. Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task. *Neuroscience letters*, 393(1):40–44, 2006. [3](#)
- [12] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020. [3](#)
- [13] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *CVPR*, 2023. [5](#), [6](#)
- [14] Dogucan Yaman, Fevziye Irem Eyiokur, Leonard Bärmann, Seymanur Akti, Hazım Kemal Ekenel, and Alexander Waibel. Audio-visual speech representation expert for enhanced talking face video generation and evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6003–6013, 2024. [7](#)