VisualSpeaker: Visually-Guided 3D Avatar Lip Synthesis

Alexandre Symeonidis-Herzig Özge Mercanoğlu Sincan Richard Bowden
CVSSP, University of Surrey, United Kingdom
{a.symeonidisherzig, o.mercanoglusincan, r.bowden}@surrey.ac.uk

Figure 1. **VisualSpeaker results.** The generated animation from the phrase "No price too high." Ground truth video (top), the FLAME meshes predicted by our approach (middle), and the 3DGS renders driven by these meshes (bottom). Note how VisualSpeaker synthesizes lip movements that accurately and expressively articulate the input, achieved by combining geometric and perceptual supervision.

Abstract

Realistic, high-fidelity 3D facial animations are essential for expressive avatars in human-computer interaction and accessibility. Although prior methods show promising quality, their reliance on the mesh domain limits their ability to fully leverage the rapid visual innovations seen in 2D computer vision and graphics. We propose ***VisualSpeaker****, a novel method that bridges this gap using photorealistic differentiable rendering, supervised by visual speech recognition, for improved 3D facial animation. Our contribution is a perceptual lip-reading loss, derived by passing photorealistic 3D Gaussian Splatting avatar renders through a pre-trained Visual Automatic Speech Recognition model during training. Evaluation on the MEAD dataset demonstrates that VisualSpeaker improves both the standard Lip Vertex Error metric by 56.1% and the perceptual quality of the generated animations, while retaining the controllability of mesh-driven animation. This perceptual focus naturally supports accurate mouthings, essential cues that disambiguate similar manual signs in sign language avatars.*

1. Introduction

People are innately skilled at recognizing and interpreting subtle facial cues, making the task of animating photorealistic 3D head avatars a challenging task as the renders face intense visual scrutiny. Demand for such avatars is increasing in areas like telepresence, interactive media, and particularly Sign Language Production (SLP). While spoken language uses mouth movements as secondary cues, sign language relies on mouthings as primary linguistic components to disambiguate similar manual signs [32]. Therefore, perceptually accurate mouthings are vital for effective sign language communication.

Despite this, prior 3D talking head methods [6, 10, 29, 38] were predominantly guided by geometry, relying on Mean Squared Error (MSE) loss over the vertices to ensure the predicted mesh vertices closely match the ground truth. While effective for reducing geometric error, these losses often yield averaged and minimally expressive animations

that fail to differentiate visemes [11], the basic visual units of speech that represent distinct mouth shapes and movements. This occurs as geometric guidance alone is disconnected from how people perceive lip movements, especially when solely reading lips. Vertex losses' inherent shortcomings have led to the explorations of additional losses, with Chae-Yeon *et al.* [4] succinctly summarizing the three criteria for perceptually accurate lip movements as temporal synchronization, expressiveness, and lip readability.

Among these criteria, we focus on lip readability, which is critical in applications such as SLP and accessibility systems for deaf and hard-of-hearing users, where visemes act as linguistic signals that disambiguate manual signs and encode grammatical markers [32]. While recent approaches [7, 9] have targeted this using mesh-level losses, such geometric proxies remain fundamentally disconnected from the end goal: generating photorealistic and intelligible renders. To bridge this gap between geometry and perception, we propose VisualSpeaker, a framework that directly optimizes for lip readability in the rendered pixel space. Our method guides an autoregressive transformer by leveraging a pretrained Visual Automatic Speech Recognition (V-ASR) model [20] to provide feedback on differentiable 3D Gaussian Splatting (3DGS) renders [14]. This approach improves fidelity while enabling novel applications, such as generating silent mouthing animations for sign language directly from text. Our main contributions are:

- A novel lip-reading loss computed directly on photorealistic 3DGS renders, closing the loop between geometric generation and perceptual evaluation.
- Evaluations demonstrating that our method surpasses the mesh-based baselines in both geometric terms and, crucially, in terms of lip-readability, validated by a user study on intelligibility.
- A text-to-mouthing application for SLP, which integrates a Text-to-Speech (TTS) system to generate accurate signing avatars from text alone, enabling scalable and audio-free avatar generation.

2. Related Work

Animating speech-driven head avatars has been a long-standing research challenge, spanning decades of work across both 2D and 3D domains [3, 6, 10, 21, 22, 42]. The field has progressed from rule-based systems to data-driven models, with growing emphasis on photorealism and perceptual quality.

Linguistic Approaches. Rules-based procedural approaches to facial animation, often following the dominance model [5], remain prevalent in production environments, such as JALI [8] and FaceFX [23]. These methods decompose speech into phonetic units, then apply hand-crafted mapping functions that transform these units into facial poses. This yields consistent, controllable animation adopted in industry, but these methods require significant linguistic and artistic expertise to adapt to new languages or identities, limiting their scalability.

Learning-based Approaches. To overcome the limitations of procedural systems, data-driven approaches learn animation directly from paired audio-visual data. Early methods [21, 22] demonstrated feasibility but were limited by small datasets and computational constraints. Subsequent methods based on RNNs and CNNs [6, 13, 29, 33] enabled more expressive animations across diverse appearances. However, even with attempts [13, 29] to encapsulate the longer-term dependencies of speech, these methods still struggled to capture the complex relationships between audio and facial movements over a broad range of identities. By adapting the Transformer architecture [35], models like FaceFormer [10] now better capture these relationships, achieving more expressive and temporally stable results by considering a longer audio context.

Despite these advances, over-smoothing remains a persistent challenge across all methods, leading to less expressive animations. This primarily stems from the reliance on L2 losses in the geometric domain, which tend to average out subtle movements. To mitigate this, works have proposed various strategies. One approach, seen in CodeTalker [38], uses a Vector Quantized-Variational AutoEncoder codebook [34] to discretize facial motions and preserve nuance through quantization. Other methods move beyond simple MSE by incorporating auxiliary losses for lip-reading, emotion, or synchronization [4, 7, 44]. While these losses improve geometric quality, they create a fundamental disconnect between the optimization target and the final, photorealistic visual output.

Photorealistic Avatars. Recent advances in novel view synthesis, particularly the real-time, differentiable rendering of 3D Gaussian Splatting (3DGS) [14], have enabled the creation of high-fidelity avatars that can be animated in real-time. This capability has opened up new approaches for 3D talking head synthesis [12, 17, 18, 39], allowing direct animation of photorealistic 3D heads. The rendering efficiency of 3DGS enables methods [18, 39] to incorporate visual losses directly during training. For example, TalkingGaussian [18] creates two motion-fields, one for the head and one for the mouth, and uses an MLP conditioned on audio features to predict Gaussian primitives to render. Training uses only visual losses on a few minutes of identity-specific data, yielding high-quality reconstructions but with limited generalization to unseen identities and no explicit control over other factors such as gaze or expressions. GaussianTalker [39], most related to our work, employs a mesh to drive the 3DGS avatar. However, it relies on mesh-based renders for computing losses in a latent lip-reading space and photometric supervision of the rendered avatars, rather than directly evaluating lip readability in the final output. This

highlights a core limitation of mesh-based pipelines: when the mesh serves only as an intermediate geometry proxy, it cannot guarantee that fine-scale lip details are preserved after neural rendering. Subtle articulatory cues such as tongue position, lip closure, and inner-mouth geometry can be smoothed out or misaligned if supervision stops at the mesh stage. This disconnect between perceptual lip accuracy and training objectives motivates the need for supervision that directly enforces lip intelligibility in the final photorealistic output, as perceived by a human observer.

3. Methodology

VisualSpeaker combines parametric 3D face modeling, differentiable photorealistic rendering, and visual speech recognition to improve lip-synchronized facial animation. This section first outlines the key components: the FLAME head model, 3D Gaussian Splatting avatars, the audio-text feature extractor, and the perceptual supervision module. Then we detail the full architecture and training procedure.

3.1. Preliminaries

FLAME [19] is a widely adopted parametric 3D face model that represents head geometry using a compact set of interpretable parameters for identity ($\beta \in \mathbb{R}^{300}$), expression ($\psi \in \mathbb{R}^{100}$), and pose ($\theta \in \mathbb{R}^{6}$). As a statistical 3D Morphable Model (3DMM) [2], it deforms a canonical mesh via linear blend skinning:

$$F(\beta, \theta, \psi) \rightarrow (\mathbf{V}, \mathbf{F}). \tag{1}$$

Here, $\mathbf{V} \in \mathbb{R}^{5142\times3}$ and $\mathbf{F} \in \mathbb{Z}^{10144\times3}$ are the vertices and faces of the FLAME mesh, respectively.

We adopt FLAME for its consistent topology and explicit factorization of static identity from dynamic expression, enabling robust mesh fitting and controllable animation. As a de facto standard, FLAME has been extended in numerous works [24, 25, 31], most relevant are works [27] that allow for rendering using 3DGS [14]. To accommodate the rendering, we rigidly attached 120 triangles representing teeth to the standard FLAME topology.

For datasets lacking 3D ground truth, we employ optimization-based tracking to obtain FLAME parameters for a 3D pseudo-ground truth. Following VHAP [26], we fit FLAME to multi-view images from MEAD [36] through a multi-stage approach. Initial stages align the mesh using landmark-based losses, while later stages incorporate photometric losses on differentiable mesh renders [28]. We add temporal regularization to minimize jitter between frames and extend VHAP to refine camera parameters initially estimated via structure-from-motion [30]. For training, we also remove all head translation and rotation, as we focus on the facial expressions. This pseudo-ground truth, while robust, has limitations. The optimization-based fitting can sometimes struggle with extreme or very rapid expressions, and its accuracy is sensitive to the initial landmark detection. This results in a 'noisier' ground truth compared to direct 3D scans , which motivates our curriculum learning strategy and the adjusted vertex weighting in later training stages.

3DGS Avatar. To generate photorealistic 3D head avatars, we employ 3DGS [14], which models appearance as a set of 3D Gaussian primitives with explicit positions, shapes, and view-dependent spherical harmonics. Unlike implicit NeRF, which typically have high computational costs for rendering, 3DGS supports differentiable, high-fidelity rendering at interactive rates. This efficiency is critical for our method, making it computationally feasible to incorporate the perceptual lip-reading loss directly into the training loop without prohibitive overhead.

We bind the 3DGS primitives to a FLAME mesh following [27], allowing photorealistic rendering controlled with mesh deformations. The fitting is optimization based and results in per-subject avatars, G. 3DGS's explicit nature enables differentiable rendering at interactive speeds. Optimization produces accurate results, but is slow and depends on varied input views and sequences. If these limitations were a concern, feed-forward approaches [16, 40, 43] could be used at minimal cost to final output quality.

During training, we render frames by passing the predicted FLAME mesh $(\mathbf{V}, \mathbf{F})$, precomputed Gaussian parameters G, and camera parameters C to the differentiable 3DGS renderer [14] producing an image $\mathbf{I} \in \mathbb{R}^{96\times96\times3}$:

$$R(\mathbf{V}, \mathbf{F}, G, C) \rightarrow \mathbf{I} \tag{2}$$

3.2. Perceptual Supervision

We introduce a novel perceptual supervision signal that evaluates lip-readability directly on photorealistic outputs, inspired by recent advances in perceptual loss for facial animation [7, 44]. Unlike prior work that computes losses on intermediate mesh representations, our method leverages an

(a) Ground Truth (b) 3DGS Render (c) Mesh Render

Figure 2. **Lip Region Comparison.** Visual comparison of lip regions after alignment, cropping, and grayscale conversion for lip-reading supervision. The 3DGS render (middle) closely resembles the ground truth (left), while the mesh render (right) lacks photorealistic detail.

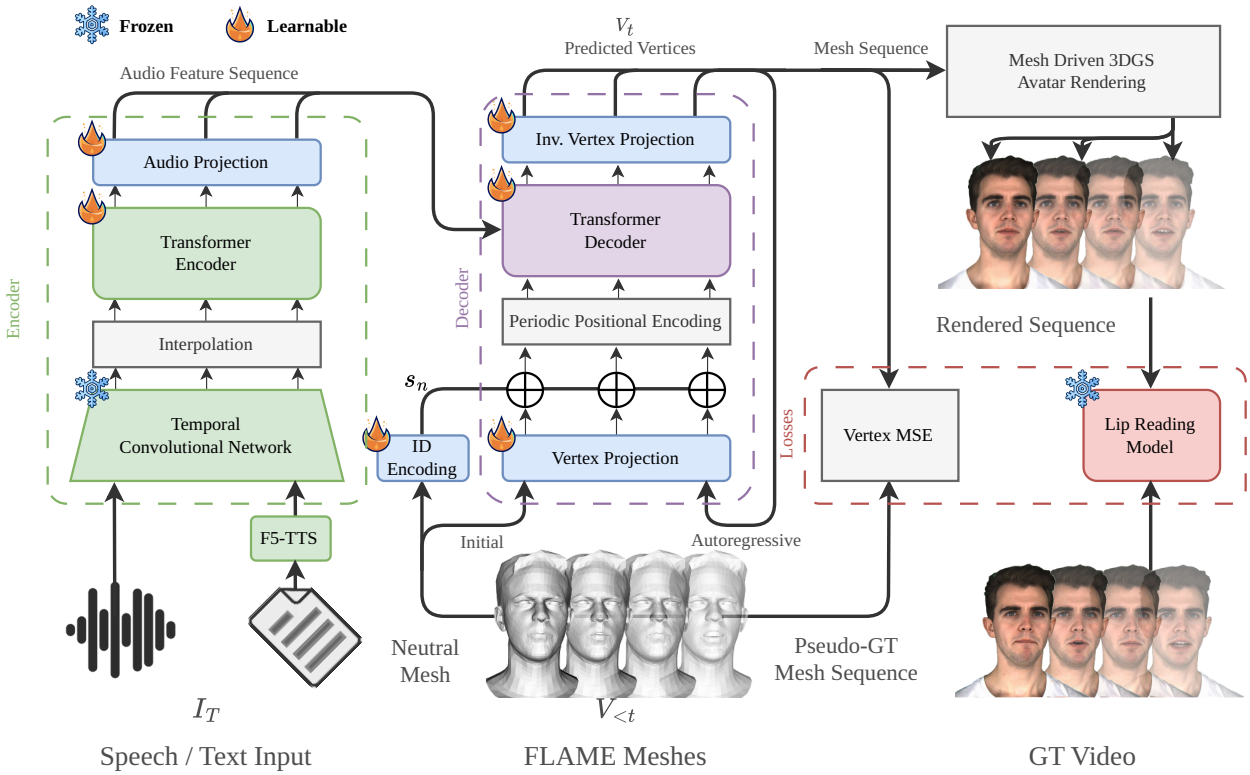

Figure 3. **Overview of VisualSpeaker.** Our encoder–decoder framework predicts the next frame's vertex offsets, $\mathbf{V}_t$. Given either text or audio, the encoder (left) generates input features I_T. These, together with past facial motion, $\mathbf{V}_{<t}$, and a speaker identity embedding, s_n, derived from a neutral FLAME mesh, are passed to the decoder (middle). During training (right), predictions are supervised by a standard vertex loss and a novel perceptual loss computed on photorealistic 3DGS renders using a pretrained lip-reading model [20].

efficient 3DGS pipeline to supervise the final visual output. This strategy closes the gap between geometric accuracy and perceptual intelligibility. As illustrated in Figure 2, our 3DGS renders achieve a visual fidelity far closer to the ground truth than mesh renders, making them a more effective target for perceptual evaluation.

To implement this, we employ the pretrained AutoAVSR model [20] to extract visual speech features, establishing a direct optimization path from the predicted FLAME mesh to the animation's perceptual quality. For computational efficiency, we render only the 96x96 pixel lip region, which is isolated using a virtual camera positioned via reprojected 3D landmarks.

A potential concern is the domain gap between synthetic 3DGS renders and real videos. To validate that our renders serve as a suitable proxy for training, we measured feature similarity within AutoAVSR's embedding space. We computed the per-frame cosine similarity between features from a ground-truth video and our render of the corresponding ground-truth mesh. Across our subset of the MEAD dataset, the matching pairs achieve a high mean cosine similarity of 0.697, while mismatched pairs fall to 0.190. These results demonstrate that our 3DGS renders are well-aligned with the AVSR embedding space, validating their suitability as supervision targets. See supplementary material for full confusion matrix.

3.3. Architecture

VisualSpeaker employs an encoder-decoder transformer architecture, drawing inspiration from Faceformer [10], as illustrated in Figure 3. The model is designed to generate 3D facial animations from either audio or text input.

Encoder. The encoder's role is to process the input modality, audio or text, and produce a sequence of feature embeddings aligned with the video frame rate.

During training, or audio-driven inference, the model processes input waveforms using a pretrained Wav2Vec2.0 model [1]. An initial Temporal Convolutional Network (TCN) extracts low-level features, which are then temporally interpolated to match the 30 FPS video frame rate, ensuring synchronization between the audio and visual streams. These features are passed through Wav2Vec2.0's transformer encoder to capture long-range contextual dependencies. To preserve the powerful, generalized audio knowledge, the TCN's weights are frozen, while the subsequent transformer layers are trained end-to-end to adapt

them to the facial animation task.

To enable direct text-to-mouthing synthesis, the framework integrates a pretrained F5-TTS model. During inference, input text is first converted into a synthetic audio waveform by the TTS model. This waveform is then processed by the same Wav2Vec2.0 encoder, creating a seamless pipeline from text to animation.

Regardless of input modality, the encoder's final output is a sequence of feature embeddings, $I_T = (i_1, ..., i_T)$, which are the cross-modal input given to our decoder.

Decoder. The decoder autoregressively predicts vertex offsets that deform a neutral FLAME mesh to create the final animation. At each step, the model predicts the next vertex offset, $\hat{\mathbf{V}}_t$, given the past offsets $\hat{\mathbf{V}}_{<t}$, the speaker embedding s_n, and the input features I_T:

$$\text{Model}(\hat{\mathbf{V}}_{<t}, s_n, I_T) \rightarrow (\hat{\mathbf{V}}_t) \tag{3}$$

The per-subject neutral mesh serves a dual purpose: it provides the base geometry to which offsets are added, and it acts as an identity prior. Its vertices are passed through a linear layer to produce a speaker embedding, s_n, which conditions the decoder to generate subject-specific morphology and articulation styles by fusing it to the projected past offsets.

During training, the sequence of past vertex offsets is projected into a 64-dimensional embedding space and fused with the speaker embedding and periodic positional encodings. This combined sequence is processed by a single transformer decoder layer composed of self-attention with a temporal bias, cross-attention with an alignment bias, and a feed-forward network. The decoder uses four attention heads and a dropout rate of 0.3. The resulting output is mapped back to the vertex offset space and added to the neutral mesh to produce the final animated mesh, $\mathbf{V}_t$. For stable training, we employ teacher forcing and a fixed learning rate, as in Faceformer [10].

3.4. Supervision

We train our model using a three-stage curriculum that strategically combines a standard geometric loss with our novel perceptual supervision. This approach avoids instability caused by applying the complex perceptual loss from the outset before a reasonable audio-to-geometry mapping is learned. Furthermore, it manages the comparatively large computational cost of differentiable rendering by introducing it only in the final refinement stage.

Our first stage is geometric pretraining, by using VOCASET [6], a dataset with high-quality 3D ground truth. This stage is supervised only by geometric vertex loss, $\mathcal{L}_{\text{vert}}$, to learn the mapping between input and facial movements. The second stage is the transition to MEAD [36] to adapt the model to the changes between the high quality 3D reconstruction and the pseudo-ground truth generated meshes. We continue to train with only $\mathcal{L}_{\text{vert}}$, but adjust weighting for focus on the lip region and account for noise in the pseudo-GT data. Finally, we introduce the novel lip-reading loss $\mathcal{L}_{\text{read}}$ and fine-tune the model on MEAD with a combined loss function.

The primary geometric loss term, $\mathcal{L}_{\text{vert}}$, is a standard vertex loss that minimizes the distance between the predicted mesh vertices and the pseudo-ground truth mesh vertices. It is calculated as a weighted MSE:

$$\mathcal{L}_{\text{vert}} = \sum_{t=1}^{T} \sum_{v=1}^{V} (||\hat{\mathbf{V}}_{t,v} - \mathbf{V}_{t,v}||^2) W_v, \tag{4}$$

where V is the total number of vertices, 5143, of the FLAME mesh with added teeth, and W_v is a per-vertex weight. During the first stage, we set W_v to 1.0 for all vertices. In the second and third stages, we reduce the weight for all non-skin vertices to 0.5 to reduce the influence of noisy pseudo-ground truth in those areas, and to 0.0 for eye vertices to ignore irrelevant reading motions present in the dataset.

Our novel perceptual loss, $\mathcal{L}_{\text{read}}$, directly evaluates the visual quality of the mouth motion in the rendered image space. First, we use the differentiable renderer to generate a sequence of lip-region images, $\hat{\mathbf{I}}_T$, from the predicted mesh sequence. Then we use a pre-trained AutoAVSR [20] to produce lip-reading features for both the predicted sequences, $\hat{\mathbf{I}}_T$, and the input frames, $\mathbf{I}_T$. The loss is the cosine distance between these feature embeddings:

$$\mathcal{L}_{\text{read}} = 1 - \text{CosSim}(\text{AutoAVSR}(\mathbf{I}_T), \text{AutoAVSR}(\hat{\mathbf{I}}_T)). \tag{5}$$

To reduce computational cost, the features from the ground truth frames are precomputed and stored.

Lip-reading supervision is applied only during the third stage of training. The total loss function is a weighted sum of the geometric and perceptual losses:

$$\mathcal{L} = \mathcal{L}_{\text{vert}} + \lambda_{\text{read}} \mathcal{L}_{\text{read}}, \tag{6}$$

where $\lambda_{\text{read}} = 1e-5$. This value was chosen empirically to scale the magnitude of $\mathcal{L}_{\text{read}}$ to be comparable to that of $\mathcal{L}_{\text{vert}}$ during training, ensuring a balanced contribution from both supervision signals. Lower values, such as $1e-6$, yielded negligible improvements over the baseline, while higher values caused artifacts, including mesh protrusions and sharp angles.

4. Experiments

To evaluate the effectiveness of VisualSpeaker, particularly the contribution of the 3DGS-based lip-reading loss, we conducted a series of experiments using the VOCASET [6] and MEAD [36] datasets. Our primary analysis focuses on

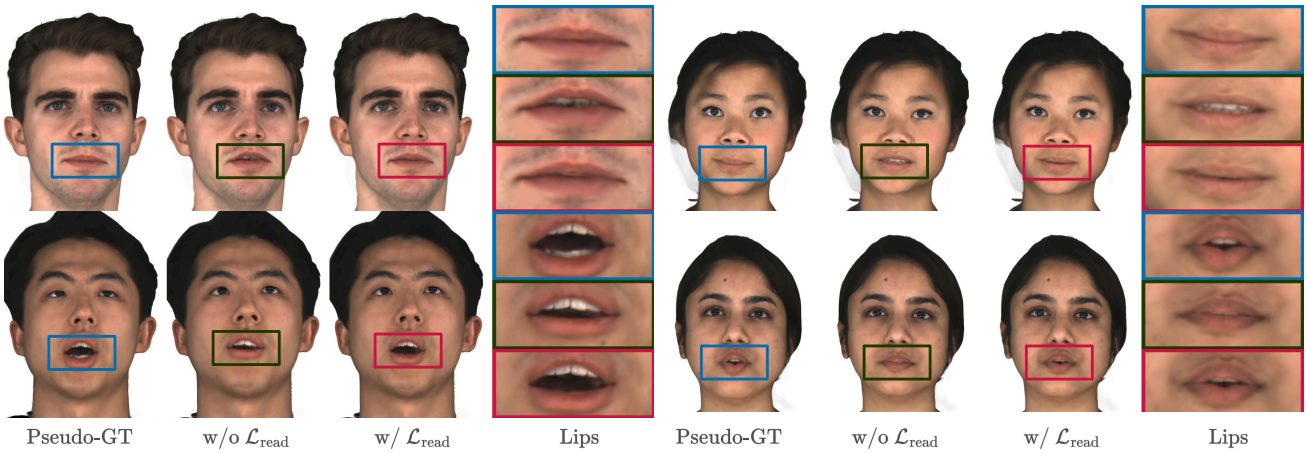

Figure 4. **Qualitative Results.** Visual comparisons for four unseen subjects and sentences from MEAD [36], highlighting how VisualSpeaker better preserves lip articulation than the baseline. Each subfigure displays three frames, left to right: Pseudo-GT render, VisualSpeaker without lip-reading loss ($\mathcal{L}_{\text{read}}$), and VisualSpeaker with full supervision. Next to these, we show a zoomed-in crop of the mouth region in the same order top to bottom, highlighting differences in lip articulation.

an ablation study comparing our full model against a variant trained without this perceptual supervision. The results demonstrate that incorporating a 3DGS-based lip-reading loss significantly improves lip-motion accuracy and perceived realism over a strong baseline, without degrading overall visual quality.

All models were implemented in PyTorch and trained on an NVIDIA 3090, with a constant learning rate of $1e-4$ and the Adam optimizer [15]. The initial stages focusing on mesh supervision were trained for 250 epochs with a batch size of four. The final stage, incorporating the lip-reading supervision, was subsequently trained for 100 epochs using a batch size of one with gradient accumulation over four steps due to memory constraints.

4.1. Datasets and Preprocessing

For initial geometric pre-training, we use VOCASET [6], which provides 480 sequences of high-quality 3D mesh data and aligned audio for 12 subjects. We follow the standard 8/2/2 subject split for training, validation, and testing, ensuring no overlap of subjects or sentences exists. Additionally, the provided 60 FPS mesh data is resampled to 30 FPS for consistency.

As VOCASET lacks any video data, we opt to use the multi-view audiovisual dataset MEAD [36] to create our photorealistic head avatars and test our lip-reading loss. We use the 40 neutral emotion sequences of the 48 actors provided.

To create the pseudo-ground truth FLAME meshes, we fit the FLAME model to the video data using the methodology outlined in Section 3.1. Due to tracking and masking challenges, five subjects were excluded. For fair evaluation, we split into six validation and test subjects with both sets containing two female and four male. As all 40 sentences are spoken by all subjects, we train only on the first 30, leaving the remaining 10 sentences to be used across the validation and test subject sets.

4.2. Quantitative Results

To evaluate the geometric accuracy of the lip movements generated by our model, we employ the standard Lip Vertex Error (LVE) metric [29]. LVE computes the maximum per-frame L2 distance between predicted and ground truth lip vertices, averaged across the sequence and is an key metric for the geometric accuracy of the animation. The average LVE scores for different stages of our pipeline, evaluated on the test sets of VOCASET and MEAD, are presented in Table 1.

Stage / Method	$\downarrow$ **LVE$_{\text{VOCASET}}$**	$\downarrow$ **LVE$_{\text{MEAD}}$**
Pretraining	3.06	7.05
VisualSpeaker w/o $\mathcal{L}_{\text{read}}$	–	3.85
VisualSpeaker	**–**	**1.69**

Table 1. **Comparison of Loss Vertex Error** (LVE (mm), lower is better) across pipeline stages. $\mathcal{L}_{\text{read}}$ represents the lip-reading loss.

The inherent differences between the datasets and the challenge posed by MEAD are immediately apparent. The model pretrained on VOCASET achieves an LVE of $3.06\,\text{mm}$ on its native test set. However, when this same model is applied to the MEAD test set, the LVE significantly increases to $7.05\,\text{mm}$, a rise of 130%. This likely reflects differences in capture conditions, with MEAD's pseudo-ground truth being noisier or more variable than

Stage / Method	PSNR ↑	SSIM ↑	LPIPS ↓
Pseudo-GT Vertices	*20.47*	*0.9126*	*0.1265*
Pretraining	**19.48**	0.9057	0.1353
VisualSpeaker w/o $\mathcal{L}_{\text{read}}$	19.29	0.9077	0.1326
VisualSpeaker	19.32	**0.9083**	**0.1316**

Table 2. **Visual Results.** PSNR, SSIM, and LPIPS scores for different training stages on the MEAD test set.

VOCASET's direct 3D scans, possibly due to tracking intricacies or less controlled recording conditions. Adapting to MEAD by fine-tuning mitigates this, improving the LVE on MEAD to $3.85\,\text{mm}$ and yet remaining 26% higher than the retrained models performance on VOCASET.

Incorporating $\mathcal{L}_{\text{read}}$ in the full VisualSpeaker model further reduces the MEAD LVE to $1.69\,\text{mm}$, representing a 56.1% improvement over the fine-tuned model and a 44.8% reduction relative to the pretrained baseline. This demonstrates that perceptual supervision from lip-reading not only improves alignment with visual intelligibility but also drives the model toward more precise geometric articulation, likely by enhancing attention to the mouth region during training.

We also assess the visual fidelity of the generated avatars with common visual quality metrics, Peak Signal-to-Noise (PSNR) Structural Similarity Index Measure (SSIM) [37], and Learned Perceptual Image Patch Similarity (LPIPS) [41], on the rendered frames. We include: (i) ground truth images, (ii) images rendered from pseudo-ground truth vertices, and (iii) outputs from successive stages of our method. Since our method does not modify the underlying 3DGS representation, the pseudo-ground truth renders serve as an upper bound for achievable image quality. Table 2 shows the results of these metrics as averages of per-sequence scores on the MEAD test set.

As shown in Table 2, improvements in these metrics across training stages are modest. The full VisualSpeaker model yields a slight gain in SSIM and LPIPS, though PSNR decreases slightly compared to the pretrained model. This discrepancy likely arises because these metrics are more sensitive to global image fidelity than to the local articulatory details, such as the lips, that our method explicitly targets. Figure 5 illustrates this, showing that while both the VisualSpeaker and pretraining models receive similar PSNR scores, the renders from the pretrained model appear visually unnatural to humans, with distorted expressions, while VisualSpeaker produces more realistic facial dynamics. These results indicate that our pipeline preserves visual quality relative to the pseudo-ground truth ceiling, and that introducing perceptual lip-reading supervision does not degrade image quality, despite not being directly optimized for pixel fidelity.

(a) Ground Truth (b) VisualSpeaker (c) Pretrained

Figure 5. **Qualitative Comparison.** Example outputs from VisualSpeaker and the model pretrained only on VOCASET, on unseen subjects and sentences from MEAD [36]. Despite the clear perceptual differences, the PSNR values for these frames are 21.11 dB and 20.90 dB, respectively.

4.3. Qualitative Results and User Study

Quantitative metrics, both geometric and visual, do not fully capture the perceptual quality of dynamic facial animations as experienced by human observers. Therefore, qualitative analysis and user studies are essential to evaluate the effectiveness of our method in generating realistic and intelligible lip movements. Figure 4 presents a series of visual comparisons between our full VisualSpeaker model, the baseline model (VisualSpeaker w/o $\mathcal{L}_{\text{read}}$), and renders from pseudo-ground truth vertices for unseen subjects and sentences from the MEAD test set. We use pseudo-ground truth vertices driven avatar, rather than real images, as a reference to isolate and evaluate the accuracy of motion synthesis, independent of texture or identity reconstruction.

Visually, the VisualSpeaker model produces animations that surpass the baseline in key aspects. For example, improved mouth closures are evident in the top row of Figure 4, while the bottom right demonstrates more expressive, large-scale lip motions. The bottom left highlights improvements in generating distinct mouth shapes, such as pursed lips. Videos of the generated animations in Figure 4 are included in the supplementary material.

To further assess the perceptual quality of our generated animations, we conducted a user study. We recruited 51 participants, who were presented with side-by-side video comparisons. Each comparison showed animations generated by VisualSpeaker versus either a baseline model (trained without lip-reading supervision), or animations rendered from pseudo-ground truth meshes. Participants were instructed to rate their preference based on visual realism of lip movements and ease of lip-reading, using a 5-point scale from *strongly prefer left* (-2), *prefer left* (-1), *no preference* (0), *prefer right* (1), to *strongly prefer right* (2). Each participant evaluated 20 randomly selected video pairs from a pool of 100 test sequences, ensuring broad coverage while maintaining manageable session length. The study interface, detailed instructions, and example comparisons inter-

Comparison Pair	Realism (%) ↑	Lip Clarity (%) ↑
Ours vs. Baseline	$63.8\% \pm 8.8\%$	$66.6\% \pm 10.5\%$
Ours vs. Pseudo-GT	$34.9\% \pm 8.0\%$	$33.6\% \pm 8.7\%$

Table 3. **User Study: Overall Preference.** Percentage of times VisualSpeaker was preferred in A/B comparisons, $\pm$ the standard deviation. 'Baseline' is VisualSpeaker w/o $\mathcal{L}_{\text{read}}$; 'Pseudo-GT' uses fitted ground truth vertices.

(a) **Why** (b) **Because** (c) *Why* (d) *Because*

Figure 6. SLP Example. Two left sub-figures show still frames of two BSL signs with identical manual and different mouthings. Right sub-figures show how VisualSpeaker can generate mouthings capable of disambiguating signs using text input.

face can be seen in the Supplementary.Preference scores were computed from these ratings to quantify perceptual advantages between methods.

The results of the user study, shown in Table 3, indicate that VisualSpeaker outperforms the baseline with 65% of participants preferring the animations generated by VisualSpeaker in terms of both realism and lip clarity. This highlights the contribution of lip-reading loss.

However, when compared to the pseudo-ground-truth vertices, our output quality still has room for improvement. This is likely due to limitations in expressiveness and a relative lack of subtle, fast lip movements in our current generation. While VisualSpeaker improves lip articulation overall, the model can under-articulate very rapid or subtle consonant closures, such as plosives (/p/, /b/) or brief tongue contacts, which the FLAME model does not explicitly capture. Additionally, users reported that eye motion, blinking, and general upper-face activity in the pseudo-ground-truth renders contributed significantly to perceived realism. This suggests that the lower gap rating for our method may be partly due to missing or under-articulated non-verbal cues beyond the mouth region, which we currently do not model. This highlights that while our method advances lip-synchronization, achieving full human-level realism in 3D avatars is a holistic challenge. Additional work to ingrate our method with systems that control upper-face expressions, eye gaze, and blinking to bridge this remaining gap is a clear direction for future work.

4.4. Sign Language Production

To evaluate our method's ability to generate linguistically meaningful mouthings, we chose sign pairs which share the same manual component in British Sign Language (BSL), differing only in their mouthings. Figure 6 shows an example of the 'why' and 'because' signs. Still frames of a real, unseen, signer are on the left and outputs from VisualSpeaker generated purely from gloss-level text input and synthesized speech are on the right. Our model successfully produces distinct mouth shapes for these two signs, capturing differences such as lip rounding and closure patterns. This is achieved using only glosses and the TTS model, without requiring paired audio or manual alignment. The complete sequences are provided in the supplementary material. The success of this application is a direct result of the improved lip readability provided by our perceptual loss, demonstrating that optimizing for visual intelligibility enables crucial downstream tasks in accessibility and human-computer interaction that are unattainable with purely geometric supervision.

5. Conclusion

We present VisualSpeaker, a method for speech-driven 3D facial animation that bridges geometric accuracy and perceptual intelligibility. By supervising directly in the rendered domain with a lip-reading loss on photorealistic differentiable 3DGS avatars, our approach achieves a 56% LVE reduction on MEAD without degrading image fidelity. A user study confirms clear gains in realism and lip clarity, and we demonstrate practical impact for text-to-mouthing in sign language, where viseme precision is essential.

On the other hand, several limitations remain. Differentiable 3DGS training incurs high computational cost and limits batch size, suggesting a need for more efficient renderers. The method also relies on per-subject multi-view avatars; future work could adapt our loss to more generalizable or few-shot avatar pipelines. Finally, our model lacks explicit control of emotion and upper-face cues, which perceptually matter for natural communication.

Overall, our results demonstrate that incorporating direct perceptual supervision at the final output level is a promising step toward more expressive and intelligible 3D avatars for accessible communication in applications such as sign language translation and human-computer interaction.

Acknowledgements

This work was supported by the SNSF project 'SMILE II' (CRSII5 193686), the Innosuisse IICT Flagship (PFFS-21-47), EPSRC grant APP24554 (SignGPT-EP/Z535370/1) and through funding from Google.org via the AI for Global Goals scheme. This work reflects only the author's views and the funders are not responsible for any use that may be made of the information it contains.

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. 4

[2] Volker Blanz and Thomas Vetter. A Morphable Model for the Synthesis of 3D Faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 187–194, 1999. 3

[3] Christoph Bregler, Michele Covell, and Malcolm Slaney. *Video Rewrite: Driving Visual Speech with Audio*, pages 353–360. 1997. 2

[4] Lee Chae-Yeon, Oh Hyun-Bin, Han EunGi, Kim Sung-Bin, Suekyeong Nam, and Tae-Hyun Oh. Perceptually Accurate 3D Talking Head Generation: New Definitions, Speech-Mesh Representation, and Evaluation Metrics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21065–21074, 2025. 2

[5] Michael M. Cohen and Dominic W. Massaro. Synthesis of a more natural-sounding talking head. *Behavior Research Methods, Instruments, & Computers*, 22(2):260–263, 1990. 2

[6] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. VOCA: Voice Operated Character Animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111, 2019. 1, 2, 5, 6

[7] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. EMOTE: Emotional Speech-Driven Animation with Content-Emotion Disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–13, 2023. 2, 3

[8] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. JALI: An Animator-Centric Viseme Model for Expressive Lip-Synchronization. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 2

[9] Han EunGi, Oh Hyun-Bin, Kim Sung-Bin, Corentin Nivelet Etcheberry, Suekyeong Nam, Janghoon Joo, and Tae-Hyun Oh. Enhancing Speech-Driven 3D Facial Animation with Audio-Visual Guidance from a Lip Reading Expert. *arXiv preprint arXiv:2407.01034*, 2024. 2

[10] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 4, 5

[11] Cletus G Fisher. Confusions among Visually Perceived Consonants. *Journal of Speech, Language, and Hearing Research*, 11(4):796–804, 1968. 2

[12] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5784–5794, 2021. 2

[13] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 2

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics (TOG)*, 42(4), 2023. 2, 3

[15] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[16] Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large Animatable Gaussian Reconstruction Model for High-Fidelity 3D Head Avatars. *arXiv preprint arXiv:2502.20220*, 2025. 3

[17] Dongze Li, Kang Zhao, Wei Wang, Yifeng Ma, Bo Peng, Yingya Zhang, and Jing Dong. S3D-NeRF: Single-Shot Speech-Driven Neural Radiance Field for High Fidelity Talking Head Synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 365–382, 2024. 2

[18] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. TalkingGaussian: Structure-Persistent 3D Talking Head Synthesis via Gaussian Splatting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 127–145, 2024. 2

[19] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a Model of Facial Shape and Expression from 4D Scans. *ACM Transactions on Graphics (TOG)*, 36(6):1–17, 2017. 3

[20] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 2, 4, 5

[21] Dominic W Massaro, Jonas Beskow, Michael M Cohen, Christopher L Fry, and Tony Rodriguez. Picture My Voice: Audio to Visual Speech Synthesis Using Artificial Neural Networks. In *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 133–138, 1999. 2

[22] Shigeo Morishima. Real-Time Talking Head Driven by Voice and Its Application to Communication and Entertainment. In *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 195–200, 1998. 2

[23] OC3 Entertainment. FaceFX, 2025. 2

[24] Foivos Paraperas Papantoniou, Panagiotis P. Filntisis, Petros Maragos, and Anastasios Roussos. Neural Emotion Director: Speech-Preserving Semantic Control of Facial Expressions in "In-the-Wild" Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18781–18790, 2022. 3

[25] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 3

[26] Shenhan Qian. VHAP: Versatile Head Alignment with Adaptive Appearance Priors. https://github.com/ShenhanQian/VHAP, 2024. 3

[27] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20299–20309, 2024. 3

[28] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3D Deep Learning with PyTorch3D. *arXiv preprint arXiv:2007.08501*, 2020. 3

[29] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. MeshTalk: 3D Face Animation from Speech Using Cross-Modality Disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1173–1182, 2021. 1, 2, 6

[30] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 3

[31] Vanessa Sklyarova, Jenya Chelishev, Andreea Dogaru, Igor Medvedev, Victor Lempitsky, and Egor Zakharov. Neural Haircut: Prior-Guided Strand-Based Hair Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19762–19773, 2023. 3

[32] Rachel Sutton-Spence and Bencie Woll. *The Linguistics of British Sign Language: An Introduction*. Cambridge University Press, 1999. 1, 2

[33] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. A Deep Learning Approach for Generalized Speech Animation. *ACM Transactions on Graphics (TOG)*, 36(4):1–11, 2017. 2

[34] Aaron Van Den Oord and Oriol Vinyals. Neural Discrete Representation Learning. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 2017. 2

[36] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3, 5, 6, 7

[37] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 7

[38] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12780–12790, 2023. 1, 2

[39] Hongyun Yu, Zhan Qu, Qihang Yu, Jianchuan Chen, Zhonghua Jiang, Zhiwen Chen, Shengyu Zhang, Jimin Xu, Fei Wu, Chengfei Lv, and Gang Yu. GaussianTalker: Speaker-Specific Talking Head Synthesis via 3D Gaussian Splatting. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM)*, pages 3548–3557, 2024. 2

[40] Dongbin Zhang, Yunfei Liu, Lijian Lin, Ye Zhu, Yang Li, Minghan Qin, Yu Li, and Haoqian Wang. GUAVA: Generalizable Upper-Body 3D Gaussian Avatars. *arXiv preprint arXiv:2505.03351*, 2025. 3

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 7

[42] 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 (CVPR)*, pages 8652–8661, 2023. 2

[43] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiyi Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan Zhang, Yongjie Zhang, et al. HeadGaP: Few-Shot 3D Head Avatar via Generalizable Gaussian Priors. *arXiv preprint arXiv:2408.06019*, 2024. 3

[44] Yixiang Zhuang, Baoping Cheng, Yao Cheng, Yuntao Jin, Renshuai Liu, Chengyang Li, Xuan Cheng, Jing Liao, and Juncong Lin. Learn2Talk: Learning to Talk and Listen from 2D and 3D Talking Faces. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2, 3