Non-verbal Audio Driven Infant Expression Synthesis

Neellohit Somayajula

May 2026

A research project report submitted to the Faculty of the Graduate School
The University at Buffalo, State University of New York

In partial fulfillment of the requirements for the degree of Master of Science

Advisor: Dr. Ifeoma Nwogu

Department of Computer Science and Engineering
University at Buffalo

Contents

Acknowledgments

I would like to thank the Department of Computer Science and Engineering at the University at Buffalo, State University of New York, and my advisor, Dr. Ifeoma Nwogu, for giving me the opportunity to work on this project under the CSE 799 Supervised Research Project. Their guidance, support, and encouragement throughout the project helped shape the direction of this work and made the research experience both meaningful and rewarding.

Abstract

The analysis of infant facial behavior is crucial for developmental research, caregiver-infant interaction studies, and early screening tools. However, raw video data of infants is highly sensitive and difficult to securely annotate or share. While generating 3D infant avatars offers a privacy-preserving alternative, existing speech-driven 3D facial animation models rely heavily on phoneme-driven articulation and adult-centric geometries, such as FLAME. Because infant vocalizations and facial dynamics are non-verbal and developmentally distinct, these adult models fail to capture accurate infant expressions.

To address this limitation, we propose a novel diffusion-based framework to synthesize 3D infant facial motion directly from non-verbal audio cues, including crying and laughing. Instead of relying on adult geometries, this approach generates facial motion within the infant-specific INFACE expression latent space. The pipeline utilizes a frozen voc2vec encoder for audio feature extraction, temporal interpolation for alignment, and a diffusion transformer equipped with banded cross-attention to handle minor audio-video misalignments. Evaluations on a curated in-the-wild benchmark dataset demonstrate that this infant-specific method substantially outperforms existing FLAME baselines, yielding significantly more realistic facial motion, lower lip geometric errors, and improved audio-lip synchronization.

Chapter 1

Motivation, Goal, and Challenges

1.1 Motivation

Infant facial behavior serves as a critical window into early human development. It provides essential signals for developmental research, allows for the study of caregiver-infant interactions, and acts as an early screening tool for potential developmental anomalies. However, working with raw video data of infants presents severe privacy and ethical challenges, making data sharing and large-scale annotation difficult. Developing accurate, privacy-preserving 3D infant avatars could resolve these issues by allowing researchers and clinicians to utilize derived 3D signals rather than identifiable video footage.

1.2 Limitations of Existing Approaches

While 3D facial animation has advanced significantly, existing models are inherently adult-centric and speech-driven. These models exhibit two primary failure modes when applied to infants:

- **Phoneme-driven articulation:** Adult speech models are trained to map distinct phonetic sounds, or phonemes, to specific visual mouth shapes. Infant vocalizations, such as crying, babbling, and laughing, are strictly non-verbal and lack this clear phoneme-viseme mapping, causing adult models to produce static or jittery outputs.
- **Adult geometry (FLAME):** The predominant 3D face model, FLAME, is derived from adult head scans. It fails to account for the unique anatomical proportions, fat distribution, and neuromuscular constraints of an infant's face, resulting in uncanny or anatomically incorrect animations.

1.3 Challenges

The scarcity of "in-the-wild" infant audio-visual datasets.

Currently, there are no large-scale, publicly available datasets that pair high-fidelity infant vocalizations, such as crying, laughing, or babbling, with corresponding 3D facial motion data. Existing infant datasets are typically heavily restricted due to privacy concerns, are strictly 2D, or focus solely on medical diagnostics rather than generative visual animation.

Consequently, training a diffusion-based model required building a custom dataset entirely from scratch by sourcing in-the-wild videos from public repositories such as YouTube and Google AudioSet.

Audio contamination and signal isolation.

Sourcing data from the wild introduced severe audio degradation issues. Videos of infants are rarely recorded in acoustically controlled settings. The raw data was heavily contaminated with background noise, including parents talking, televisions playing, ambient room echo, and background music. Because the proposed model relies on a voc2vec encoder to extract precise non-verbal audio embeddings, feeding it contaminated audio would result in jittery, inaccurate facial latents.

Visual occlusions and strict framing requirements.

Beyond the audio constraints, the visual requirements for accurate monocular reconstruction necessitated aggressive data filtering. To successfully map the 434 facial landmarks required for the INFACE latent fitting, the infant's face had to be entirely visible and facing the camera for sustained, continuous periods. In reality, infant behavior in home videos is highly erratic. Infants frequently turn their heads out of profile, and their faces are constantly occluded by parents' hands, pacifiers, bottles, or toys. A massive portion of the initially collected video pool had to be discarded because the mouth region, the most critical area for analyzing crying and laughing dynamics, was partially obscured or moved out of the frame mid-vocalization.

Chapter 2

Model Framework

Figure 2.1 summarizes the proposed training and inference pipeline. During training, infant audio cues are encoded with voc2vec-ls-pt, temporally aligned with video frames, and used to condition a diffusion transformer that predicts clean expression latents from noisy expression latents. During inference, the generated expression latents are concatenated with identity and age latents before being decoded by INFACE into a mesh sequence.

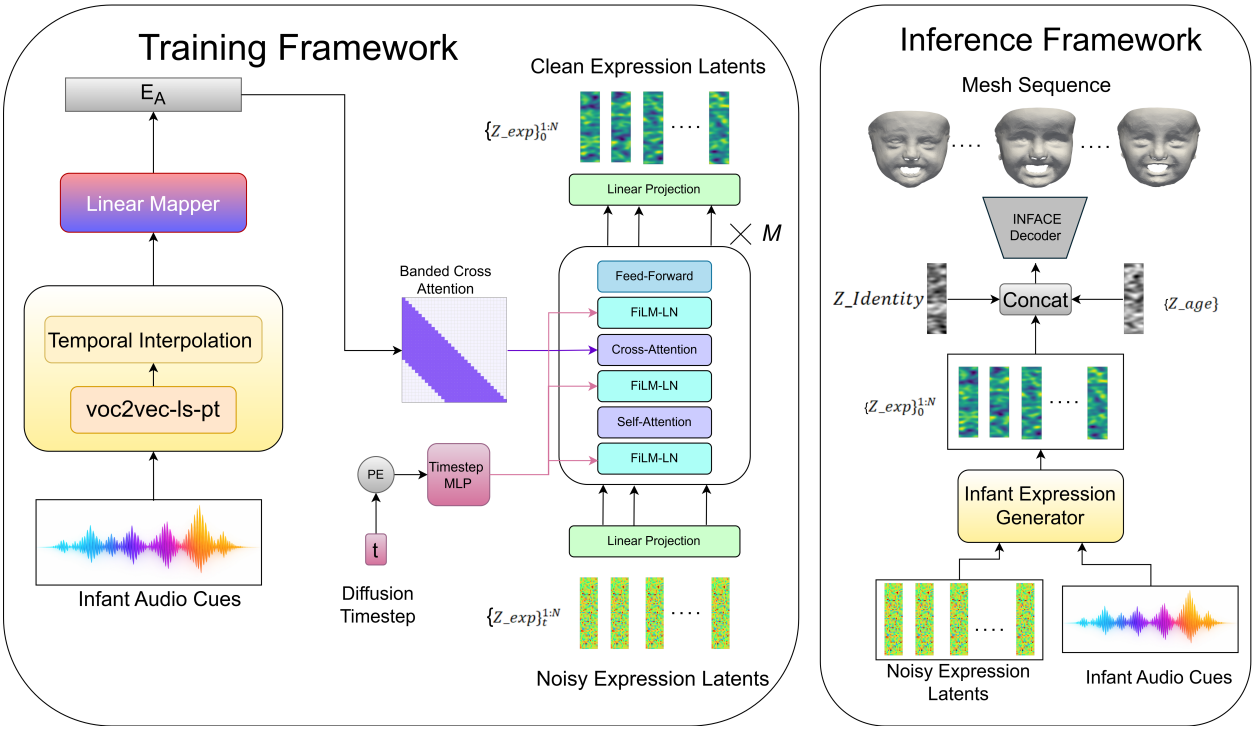

Figure 2.1: Training and inference framework for non-verbal audio-driven infant expression synthesis.

2.1 Audio Encoder and Temporal Alignment

The input non-verbal infant audio is resampled to 16 kHz. To extract rich, context-aware acoustic features without relying on speech-specific phoneme recognition, we process the raw audio using a frozen voc2vec-ls-pt encoder. Because the sampling rate of the audio embeddings inherently differs from the video frame rate, typically 30 FPS, we apply a temporal interpolation layer. This ensures

that the extracted audio latents are synchronized with the corresponding video frame timestamps.

2.2 Diffusion Generator and Attention

The core of the generative pipeline is a diffusion transformer that denoises expression latents over time. Instead of predicting the noise, the model predicts the clean expression signal directly. The architecture utilizes two distinct attention mechanisms:

- **Causal self-attention:** This ensures temporal consistency and preserves causal sequence order, preventing the model from looking ahead inappropriately, which is critical for real-time or streaming inference applications.
- **Banded cross-attention:** Non-verbal audio cues like cries and laughs often have subtle temporal misalignments with the visual mouth opening. Banded cross-attention restricts the attention map to a localized temporal window, allowing the network to handle these minor audio-video phase shifts without losing synchronization.

2.3 3D Output via INFACE Decoder

Once the clean expression latents are generated by the diffusion model, they are passed to the INFACE decoder. The dynamically predicted expression parameters are concatenated with the statically fixed identity and age parameters. The final output is a temporally smooth, anatomically accurate animated 3D infant mesh sequence.

Chapter 3

Training Objective and Losses

To ensure that the generated facial motions are not only statistically likely but also physically realistic and geometrically accurate at the lip level, the network is trained using a composite weighted objective function. The total loss encompasses several targeted terms and can be summarized as

$$\mathcal{L}_{\text{total}} = \lambda_{\text{latent}}\mathcal{L}_{\text{latent}} + \lambda_{\text{velocity}}\mathcal{L}_{\text{velocity}} + \lambda_{\text{lip}}\mathcal{L}_{\text{lip}} + \lambda_{\text{lipvel}}\mathcal{L}_{\text{lipvel}} + \lambda_{\text{DTW}}\mathcal{L}_{\text{DTW}} + \lambda_{\text{closed}}\mathcal{L}_{\text{BCE}}.$$

3.1 Latent Space Losses

The primary generative loss is the diffusion reconstruction loss, which computes the L_2 distance between the predicted clean latents and the ground-truth latents:

$$\mathcal{L}_{\text{diff}} = \frac{1}{|\Omega|} \sum_{(b,i)\in\Omega} \left\| \hat{\mathbf{x}}_{0,i}^{(b)} - \mathbf{x}_{0,i}^{(b)} \right\|_2^2 . \tag{3.1}$$

To enforce temporal smoothness and prevent inter-frame jitter, a velocity loss is applied to the first derivative, or differences between adjacent frames, of the latents.

3.2 Mesh and Geometric Losses

Because small errors in the latent space can manifest as severe geometric distortions on the final mesh, particularly around the highly expressive mouth region, the latents are decoded into meshes during training and specific lip constraints are applied:

$$\mathcal{L}_{\text{mesh-lip}} = \left\| \hat{\mathbf{V}}_{\text{lip}}^{\text{mesh}} - \mathbf{V}_{\text{lip}} \right\|_2^2 . \tag{3.2}$$

- **Mesh-decoded lip loss:** Penalizes positional errors of the lip vertices.
- **Mesh-decoded lip velocity loss:** Ensures that the speed of lip movements matches the ground truth, preventing unnaturally fast snapping motions.
- **Mouth aperture and closed losses:** Crying and laughing are characterized by wide, sustained mouth openings. Dynamic time warping (DTW) distance is penalized on the mouth aperture, and a binary cross-entropy (BCE) loss is used to ensure that the model accurately predicts when the mouth should be completely closed.

The final weighted objective balances these constraints to optimize both global expression accuracy and localized lip synchronization.

Chapter 4

Qualitative Results

Visual inspection of the generated animations reveals a stark contrast between the proposed approach and existing models:

- **Baseline limitations:** Adult-centric, FLAME-based models, including FaceFormer, Code Talker, and Imitator, consistently fail to animate infant faces accurately. They produce mouth motions that are either entirely static or only weakly changing, making them fundamentally unable to translate non-verbal cries or laughter into corresponding high-energy facial dynamics.
- **Proposed model accuracy:** In contrast, the INFACE-based diffusion model closely tracks the ground-truth mouth motion. The network successfully captures the wide jaw drops characteristic of a crying infant and the rhythmic cheek and mouth movements associated with laughter.

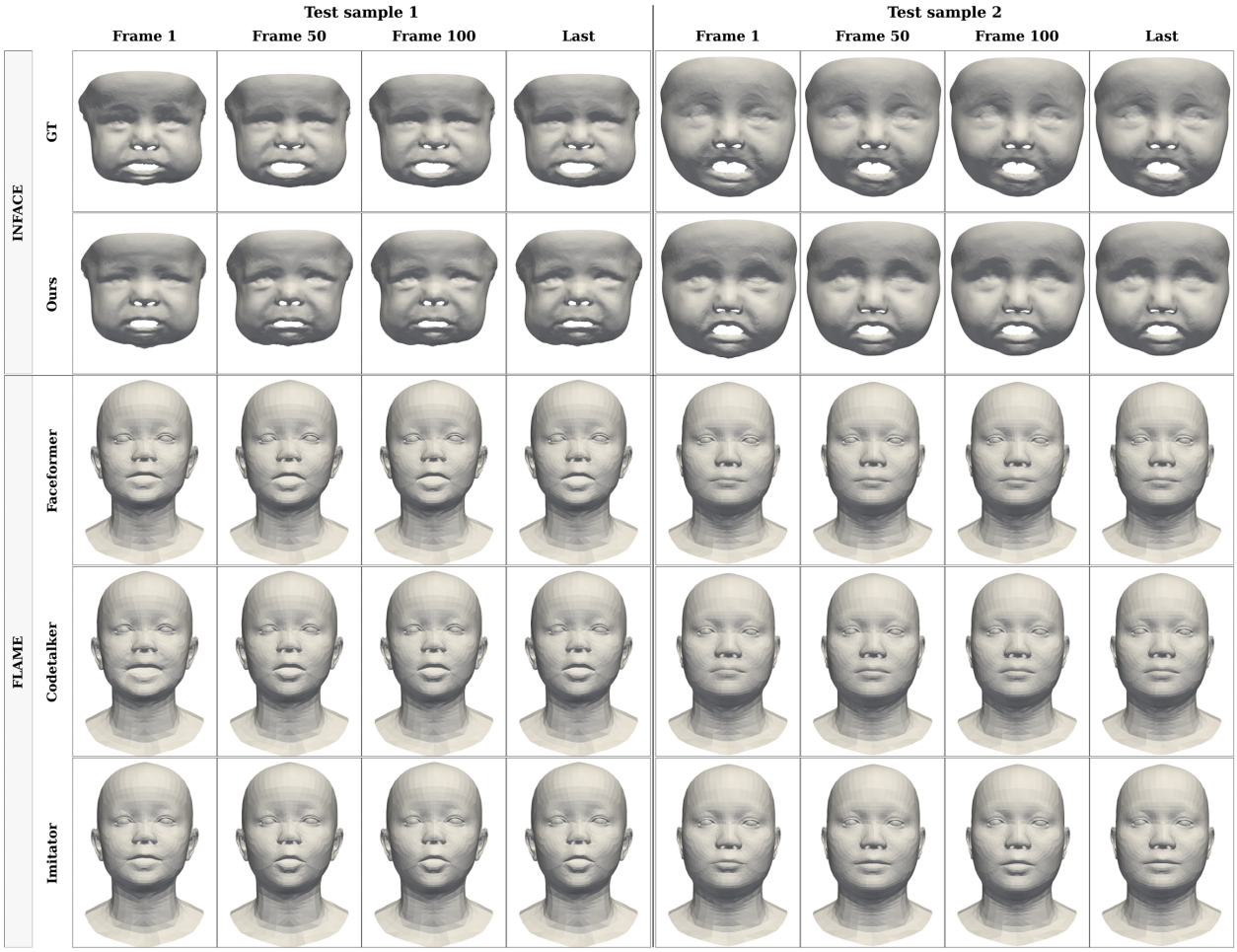

Figure 4.1: Qualitative comparison across two test samples. INFACE ground truth and the proposed method preserve infant-specific expression dynamics, while adult FLAME-based baselines show limited or inaccurate mouth motion.

Chapter 5

Evaluation Metrics

The model is quantitatively evaluated against baselines using the following specialized metrics:

- **LVE (Lip Vertex Error):** Measures the maximum L_2 geometric error across lip vertices between the predicted and ground-truth meshes. Lower values indicate better geometric accuracy.
- **DTW (Dynamic Time Warping Distance):** Evaluates the alignment between the predicted and ground-truth mouth-aperture trajectories over time, accounting for minor temporal shifts. Lower values indicate superior temporal alignment.
- **L-PCC (Pearson Correlation Coefficient):** Measures the linear correlation of lip motion over time. Higher values signify better synchronization.
- **L-CCC (Concordance Correlation Coefficient):** Assesses the agreement between the predicted and actual lip motion distributions. Higher values indicate better performance.

Chapter 6

Main Quantitative Results

The quantitative evaluation establishes the superiority of the proposed framework. As shown in Table 6.1, the proposed INFACE diffusion model achieves substantially lower Lip Vertex Error (LVE) and Dynamic Time Warping (DTW) distance than the FLAME-based baselines. It also produces stronger lip-motion correlation and concordance, as reflected by the higher L-PCC and L-CCC scores.

Table 6.1: Main quantitative results. Down arrows indicate lower is better, and up arrows indicate higher is better.

Method	LVE ↓	DTW ↓	L-PCC ↑	L-CCC ↑
Ours (INFACE diffusion) [1]	**4.518**	**0.742**	**0.111**	**0.074**
FaceFormer (FLAME) [2]	25.35	2.48	0.035	0.010
CodeTalker (FLAME) [3]	25.37	2.48	0.045	0.011
Imitator (FLAME) [4]	25.33	2.46	0.039	0.013

Chapter 7

Main Takeaways and Future Work

This project demonstrates that standardizing on adult facial geometries and speech-driven paradigms severely restricts the modeling of non-verbal, infant-specific behaviors. The main takeaways and future directions are summarized below:

- **Architectural success:** The combination of infant-specific INFACE geometry, non-verbal audio conditioning through voc2vec, and a diffusion-based generative backbone results in a robust framework for 3D infant facial motion synthesis.
- **Performance breakthrough:** The model significantly outclasses FLAME-based baselines across all geometric and temporal synchronization metrics, showing that tailored latent spaces are necessary for this domain.
- **Clinical and privacy implications:** By providing a mechanism to generate realistic 3D avatars directly from audio, this framework enables privacy-aware evaluations, allowing researchers to study developmental signals without exposing raw infant video data.
- **Future work:** Subsequent iterations of this research will focus on expanding the diversity of vocalization types, including babbling and feeding sounds, implementing stronger occlusion-handling mechanisms for in-the-wild videos, and improving robustness against extreme head poses frequently observed in infant behavior.

Bibliography

[1] L. Thies et al. INFACE: Large-Scale 3D Infant Face Model. MICCAI, 2024.

[2] X. Peng et al. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. CVPR, 2022.

[3] J. Xing et al. Code Talker: Speech-Driven 3D Facial Animation with Discrete Motion Prior. CVPR, 2023.

[4] B. Jiang et al. Imitator: Personalized Speech-driven 3D Facial Animation. ICCV, 2023.