DiffPoseTalk: Speech-Driven Stylistic 3D Facial Animation and Head Pose Generation via Diffusion Models

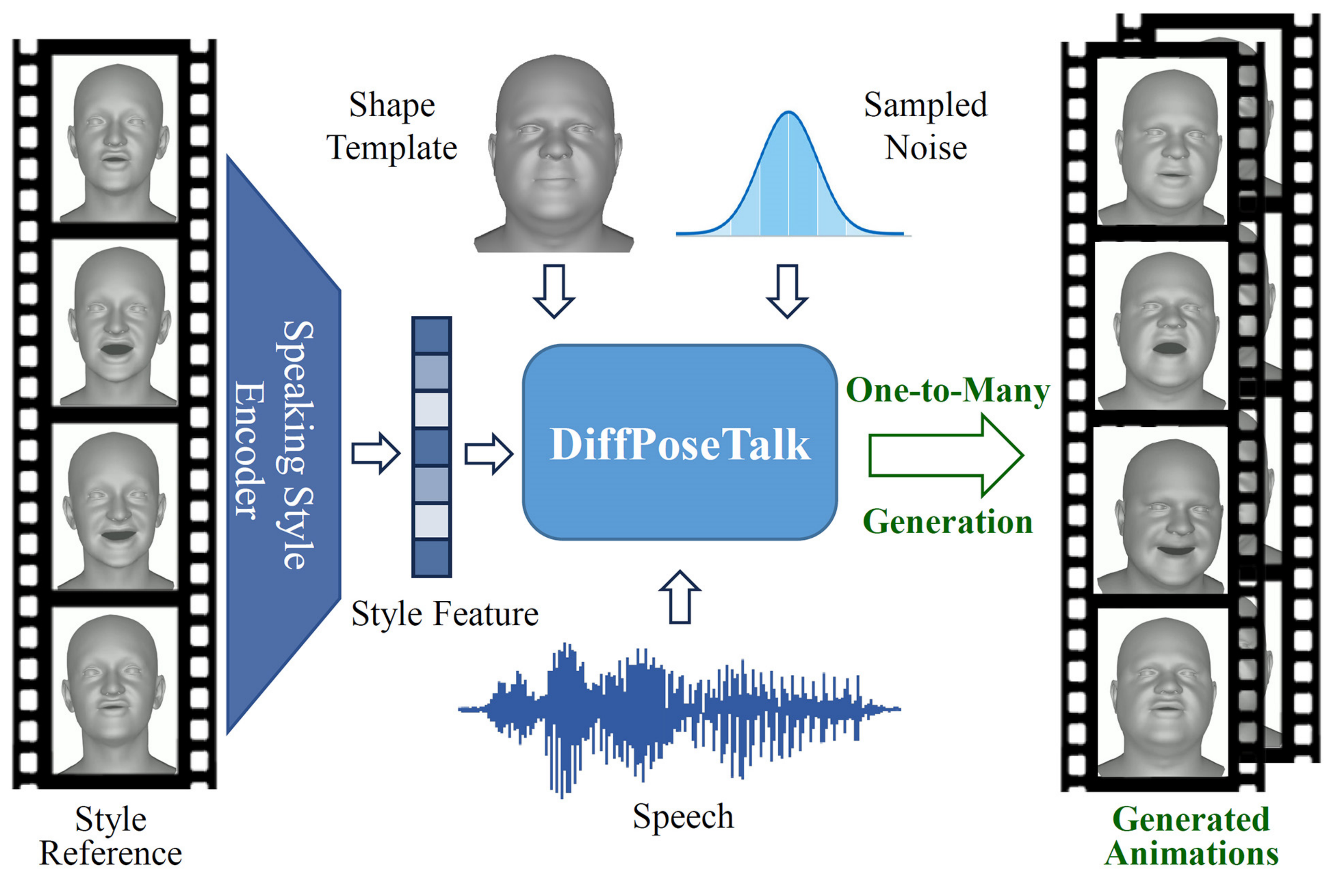
Project Page

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Zhiyao Sun[1], Tian Lv[1], Sheng Ye[1], Matthieu Lin[1], Jenny Sheng[1], Yu-Hui Wen[2], Minjing Yu[3], Yong-Jin Liu[1]

[1]*BNRist, Tsinghua University*, [2]*Beijing Jiaotong University*, [3]*Tianjin University*

Overview



DiffPoseTalk introduces a novel **diffusion-based** system for generating speech-driven facial animations and head poses, featuring **example-based style control** through contrastive learning. It overcomes the scarcity of 3D talking face data by utilizing reconstructed 3DMM parameters from a **newly developed audio-visual dataset**, enabling the generation of diverse and stylistic motions.

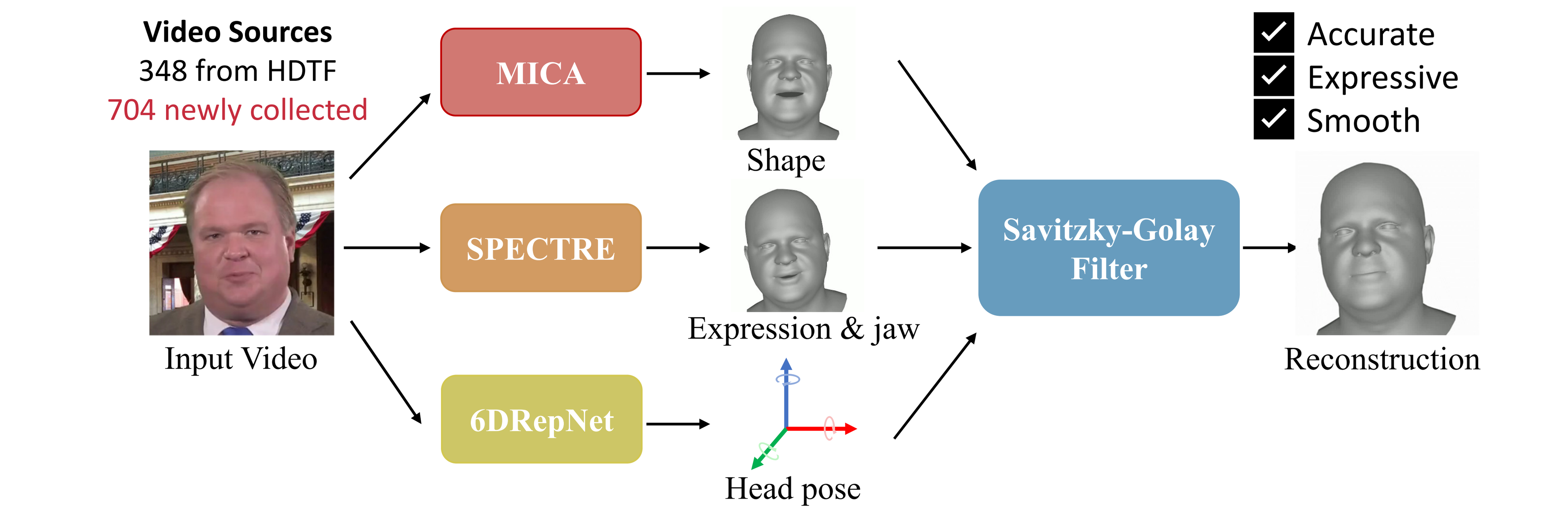
Background

Speech-driven facial animation generation has broad applications in education, entertainment, and virtual reality. However, there remain several challenges that have not been well addressed:

- **Many-to-many mapping between speech and motion**
  - A mouth shape may correspond to several sound, and vice versa. This mapping is further influenced by factors such as identity and speaking style. Non-verbal actions also exhibit a high degree of randomness.
  - Previous methods mostly model this as a **regression task**, which is not suitable enough. The generation is deterministic and suffers from the **"regression-to-mean"** problem. This also hinders the ability to generate non-verbal movements and natural head movements.
- **Style control**
  - Speaking style is a multifaceted attribute that is difficult to quantify.
  - Existing methods primarily adopt "one-hot identity labels" as style conditions, which are limited to the training set.
- **Lack of large-scale 3D facial animation dataset**
  - Collecting such data requires professional devices and is time-consuming. Thus, existing datasets have limited coverage of identities, styles, and head motions.

Method

Dataset Construction

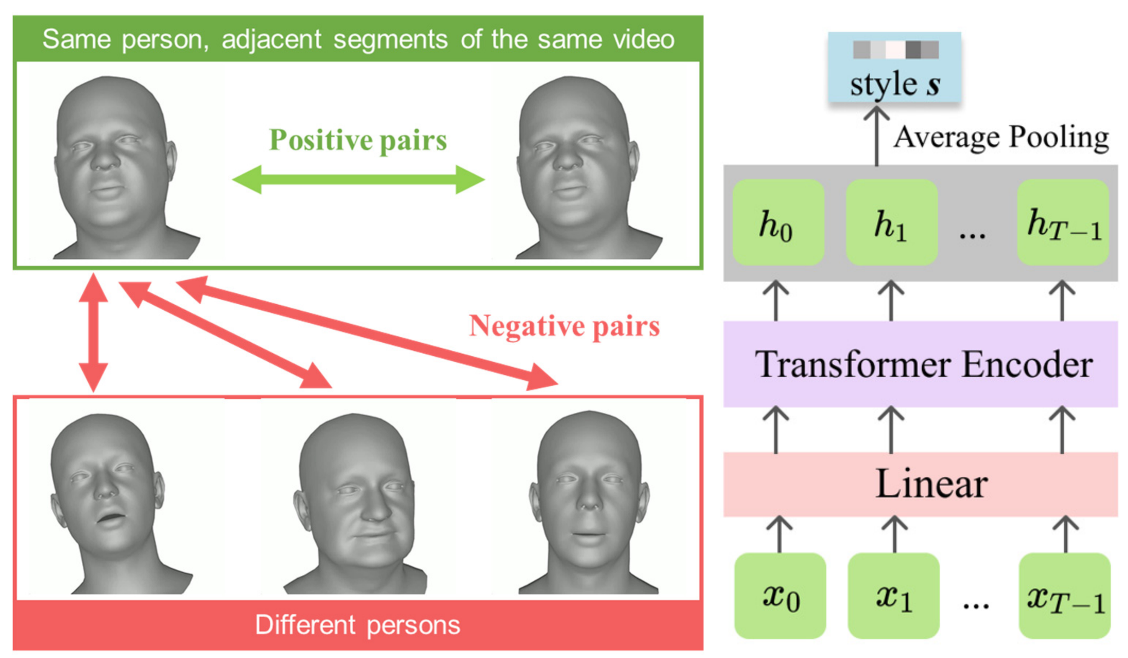


We collect a new audio-visual dataset, featuring video clips from lectures, online courses, interviews, and news programs, thereby capturing a wider array of speaking styles and head movements. State-of-the-art 3D face reconstruction and pose estimation methods are used to predict accurate, expressive, and smooth facial animations.

Speaking Style Encoder

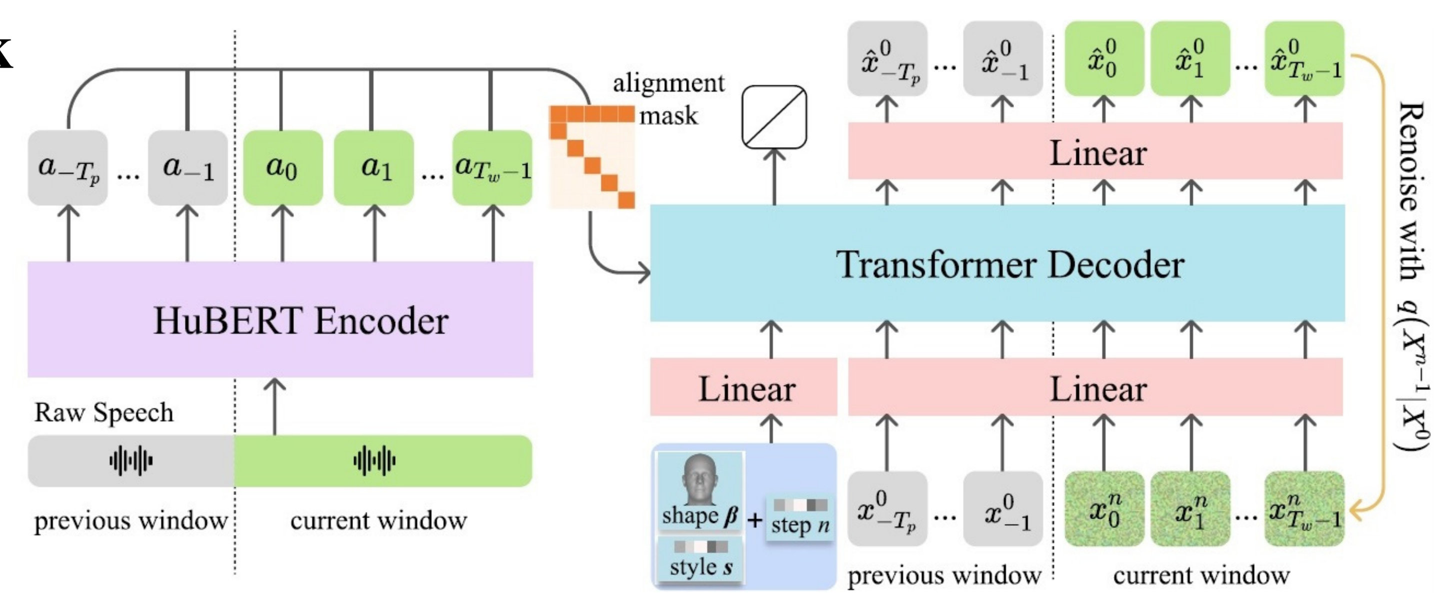
We have an important observation: the short-term speaking styles of the same person at two proximate times are often similar.

Therefore, we employ **contrastive learning** and use the InfoNCE loss to train the speaking style encoder. The encoder can extract implicit style features from any motion sequence.

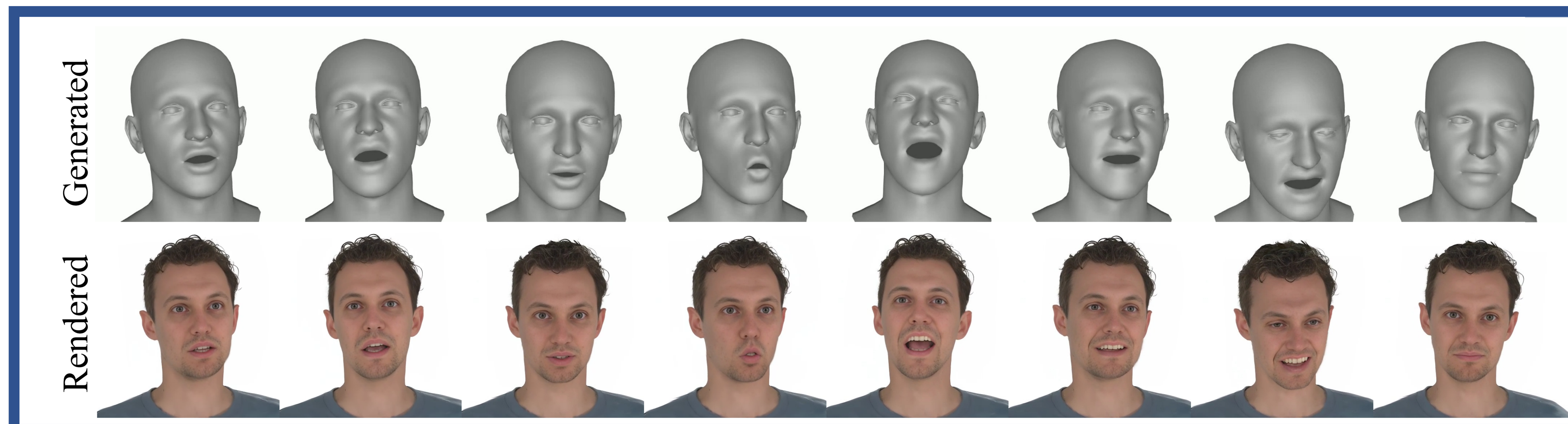


Denoising Network

- A HuBERT encoder for robust speech feature extraction
- A transformer decoder for iterative denoising

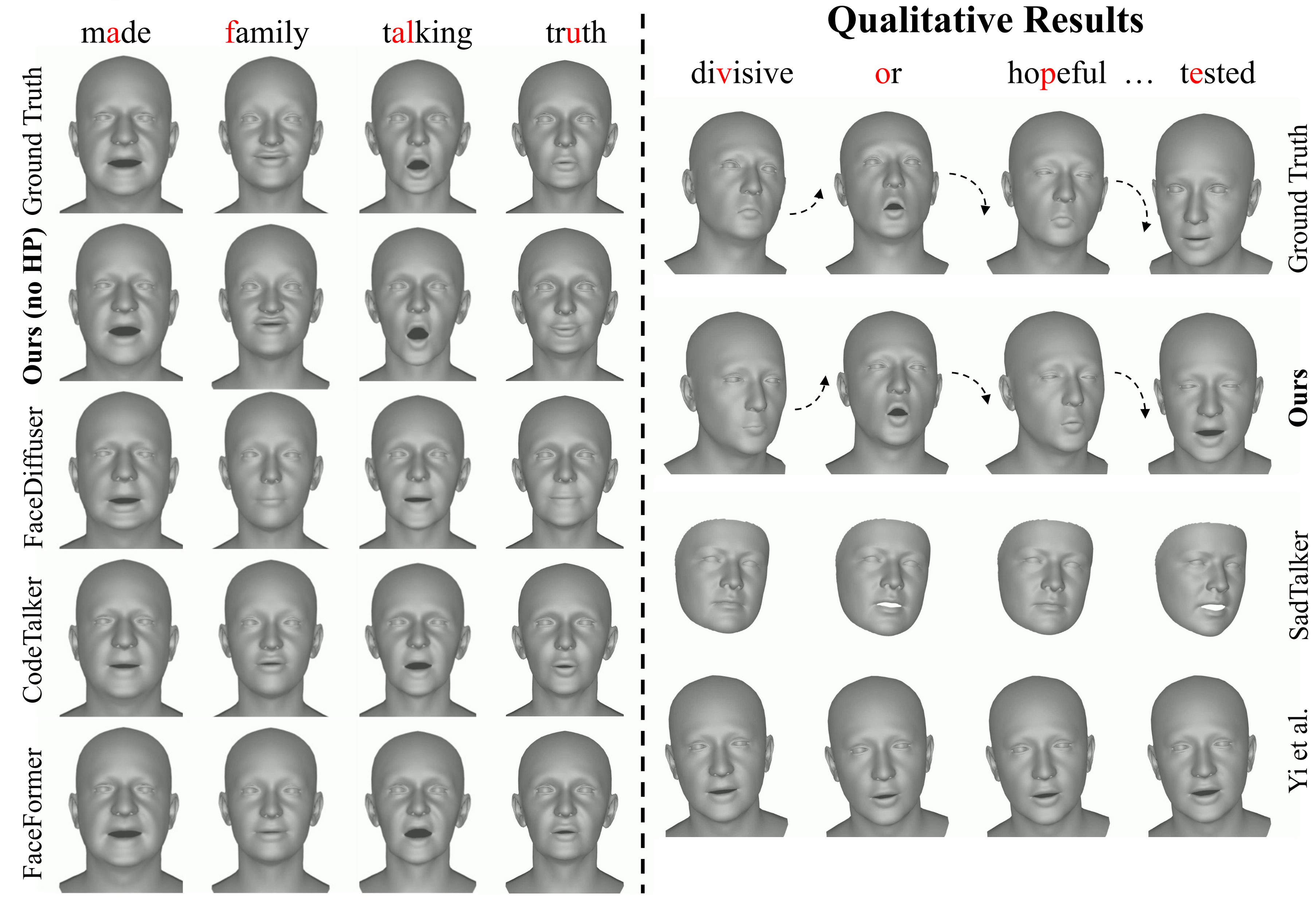


- Designs
  - Prediction of clean samples rather than noise to enable precise geometric constraints
  - An alignment mask to ensure proper alignment of the speech and motion modalities
  - A windowing strategy for arbitrary length generation



Experiments

Comparison with SoTA



Quantitative Results / **User Study**

	Methods	LVE (mm)↓	FDD ($\times 10^{-5}$m)↓	MOD (mm)↓	BA↑	Lip Sync ↑	Style Sim ↑	Natural ↑
w/o HP	FaceFormer	9.90	16.95	2.63	N/A	2.56	2.60	2.36
	CodeTalker	12.71	12.44	2.87	N/A	2.88	3.00	2.90
	FaceDiffuser	12.12	15.48	3.50	N/A	2.71	2.51	2.35
	Ours (no HP)	**8.81**	10.13	1.72	N/A	**4.23**	**4.07**	**4.43**
w/ HP	Yi et al.	9.99	21.50	2.42	0.26	1.94	2.02	1.99
	SadTalker	—	—	—	0.24	3.25	2.91	2.96
	Ours	8.94	9.60	1.62	**0.29**	**4.52**	**4.25**	**4.43**
Ablations	Ours w/o $\mathcal{L}_{geo}$	11.29	15.11	2.14	0.28			
	Ours w/o AM	12.81	12.58	2.18	0.24			
	Ours w/o CFG	9.58	**9.59**	**1.56**	**0.29**			
	Ours w/o SSE	11.33	12.97	2.03	0.28			

Please watch the demo video at the project page to see more experimental results, such as example-based style control, one-to-many generation, noisy audio, and multi-lingual results.

Discussion

- **The choice of 3DMM parameters as the face representation**
  - Reduce computational cost for the diffusion model
  - 3DMM serves as a prior to simplify training and improve generalization
  - Easier integration with downstream tasks (e.g., drive a GaussianAvatar)
- **Limitations and future work**
  - Emotion and fine-grained control.
  - Diffusion models are relatively slow. The generation speed can be further improved.
  - Modeling and animating inner mouth structure.
  - Collecting real-world 3D talking data with broader coverage and diversity.