DR^2: Disentangled Recurrent Representation Learning for Data-efficient Speech Video Synthesis

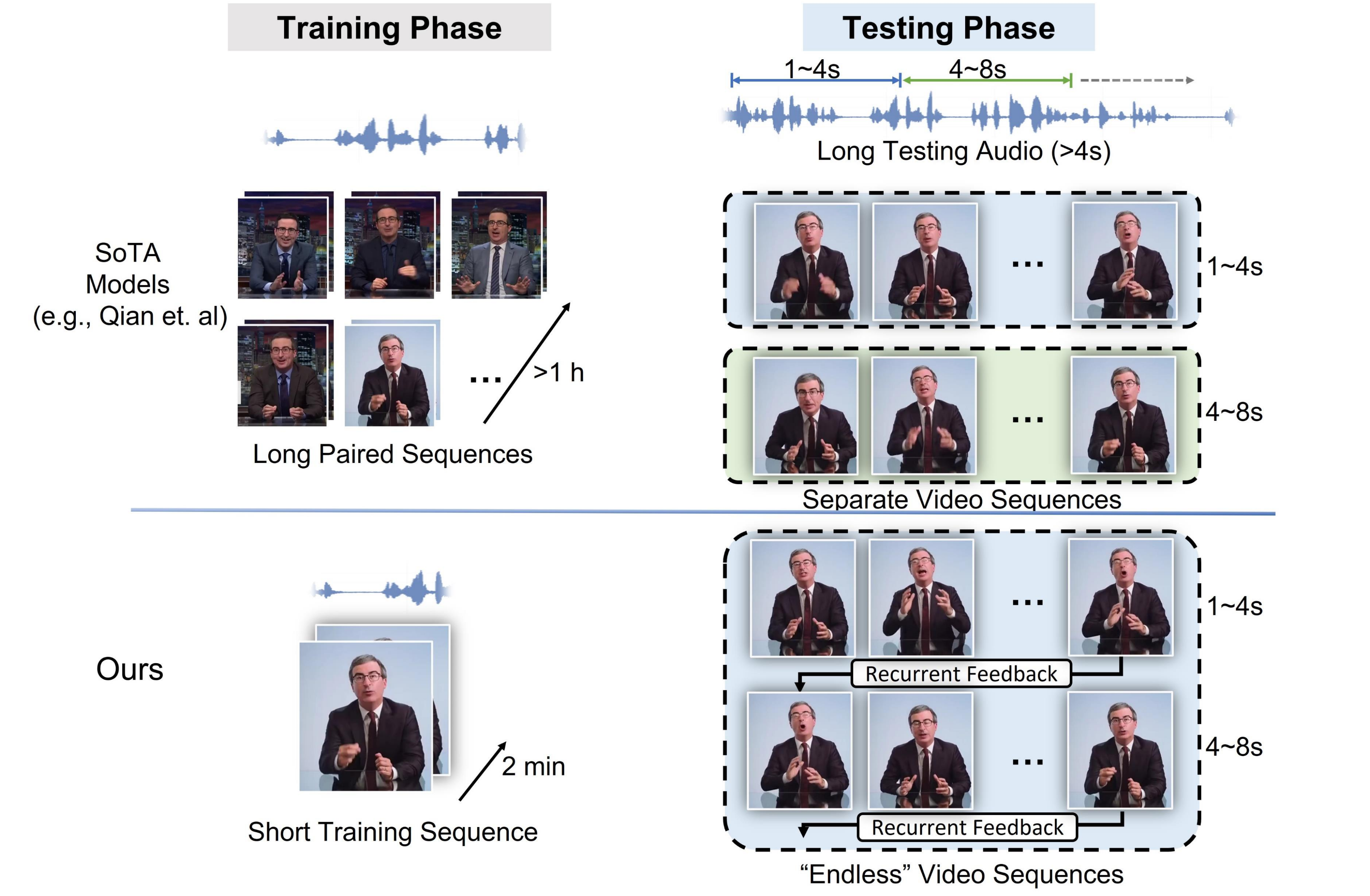
Chenxu Zhang[1], Chao Wang[1], Yifan Zhao[2], Shuo Cheng[3], Linjie Luo[1], Xiaohu Guo[4]

[1]ByteDance Inc, [2]Peking University, [3]Georgia Institute of Technology
[4]The University of Texas at Dallas

JAN 4-8 WACV 2024 WAIKOLOA HAWAII

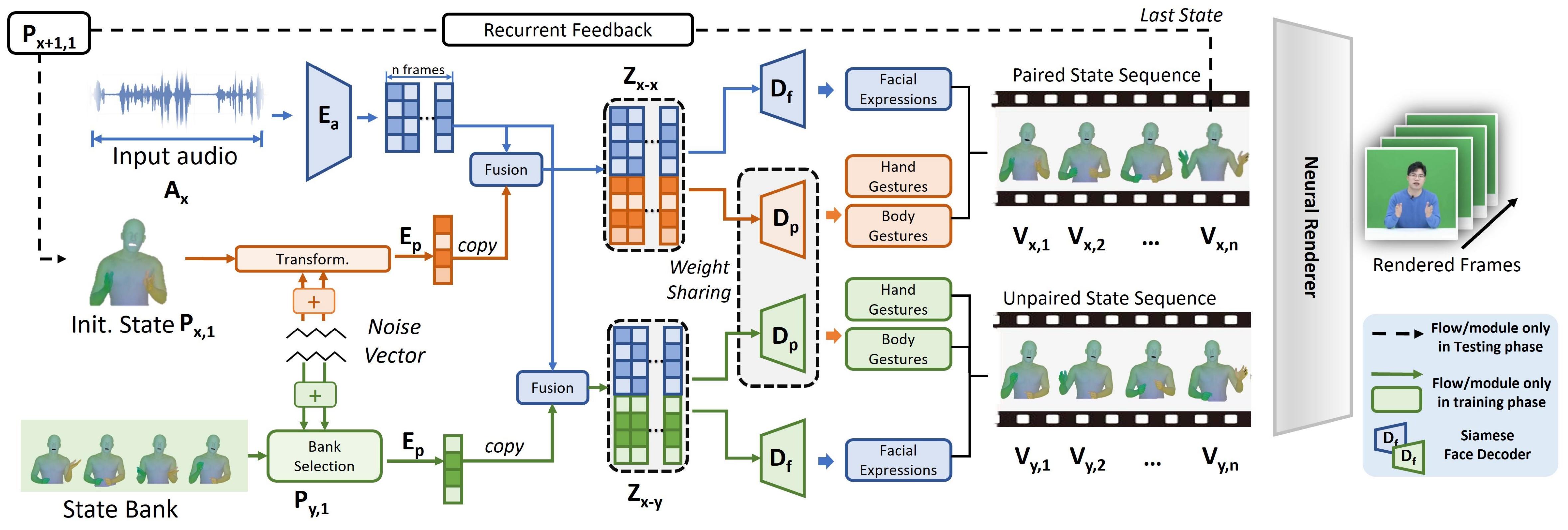
More information: https://zhangchenxu528.github.io/

Motivation



1) Training phase: using **only 2 mins short videos**.
2) Testing phase: generate **endless video sequences** with high diversity and continuity.

Method



Pipeline of our disentangled recurrent representation learning framework.

Standard Paired Training

Learning relations of input audio and pose sequences with **only initial pose**

Unpaired Training

Embedding the unpaired pose and audio **in the same latent space**

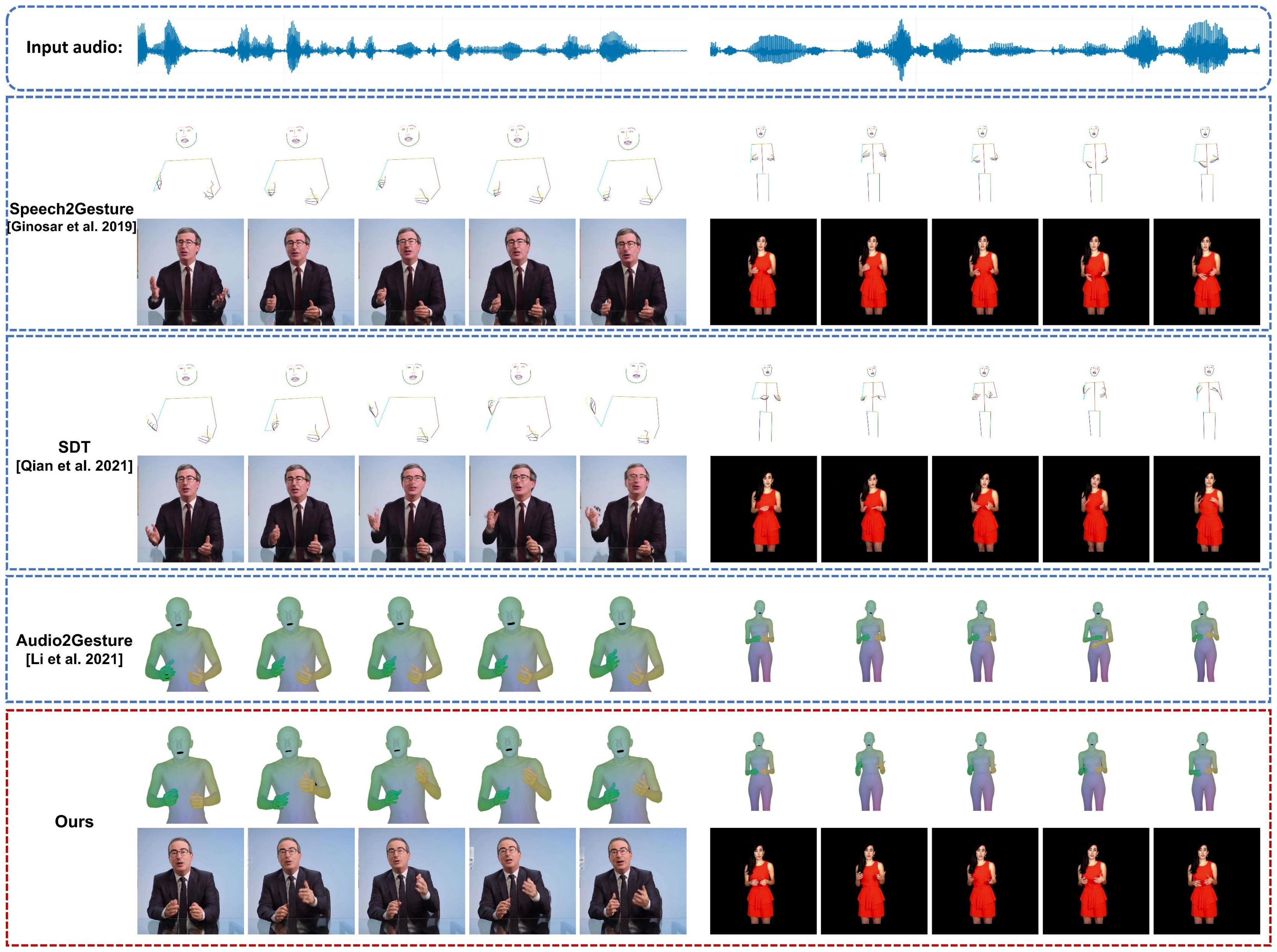
Infinite Recurrent Inference

Using last state as **recurrent guidance** to enhance ***diversity*** and ***consistency***

Gesture Neural Rendering

Rendering to realistic Videos with **high quality**

Qualitative Evaluation


Input audio:
Speech2Gesture [Ginosar et al. 2019]
SDT [Qian et al. 2021]
Audio2Gesture [Li et al. 2021]
Ours

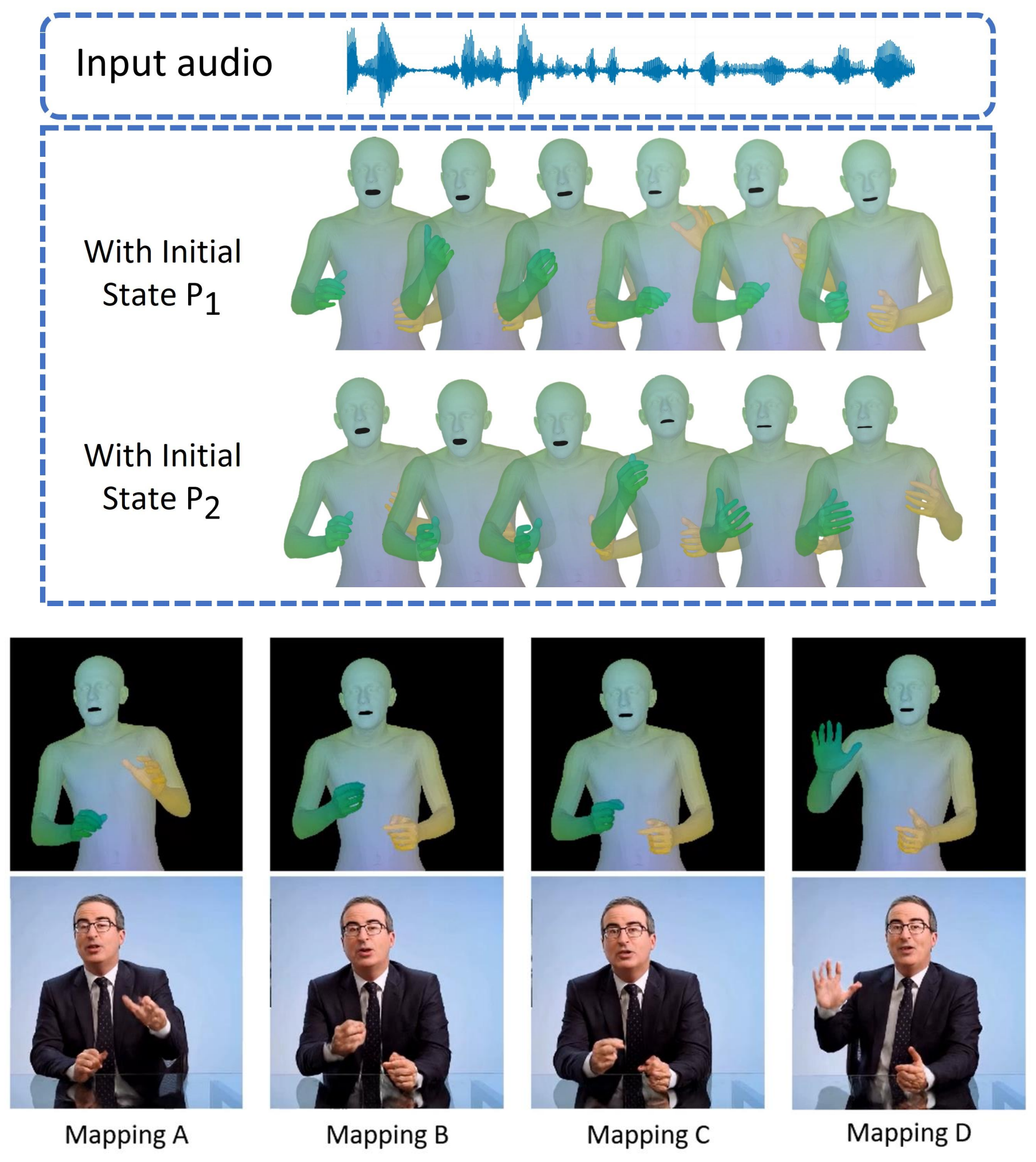
Comparisons of state-of-the-art models with our method. Note that Speech2Gesture and SDT relies on 2D skeletons as intermediate representation while Audio2gesture only generates sequences of 3D models. Compared with these works, our results show sufficient gesture diversity with varying input audios.

Quantitative Evaluation

Dataset	Method	Offset/confidence	LPIPS	CPBD	FVD	Diversity	Multimodality
Online videos	Speech2Gesture [23]	-3/1.751	0.267	0.520	636.7	6.430	-
	SDT [49]	-3/1.923	0.253	0.511	544.0	8.810	7.796
	Audio2Gesture [36]	-	-	-	-	13.647	11.747
	Ours	**-2/2.328**	**0.156**	**0.569**	**387.3**	**16.915**	**15.931**
Self-captured videos	Speech2Gesture [23]	1/0.570	0.196	0.492	310.1	8.913	-
	SDT [49]	-2/1.275	0.187	0.502	302.4	8.513	8.159
	Audio2Gesture [36]	-	-	-	-	11.318	12.553
	Ours	1/**2.029**	**0.135**	**0.526**	**246.5**	**13.447**	**16.472**

Lip-sync metric, Image quality metric, Temporal-level metric, Diversity and Multimodality metric.

One-to-Many



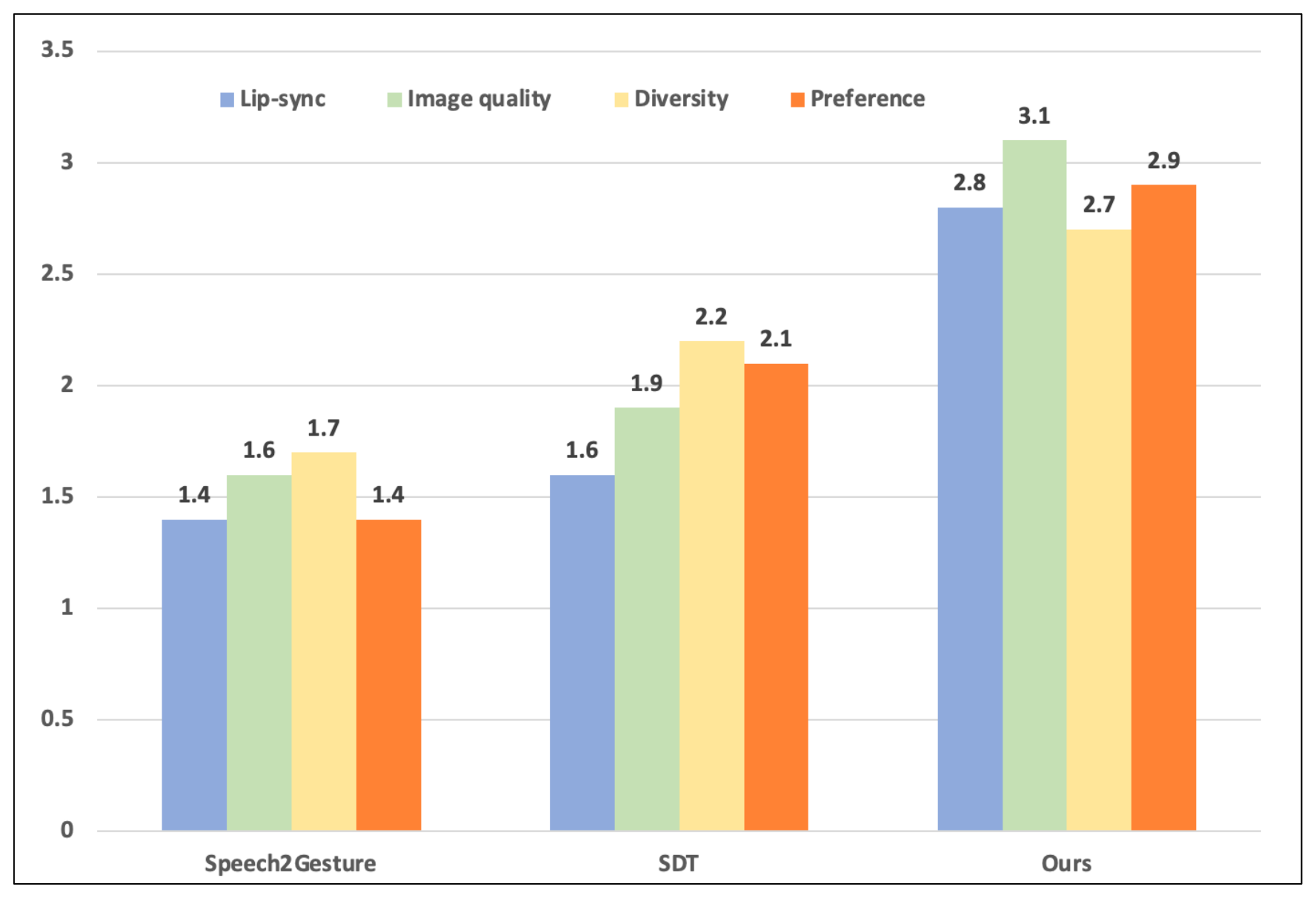
Input: one same given audio
Output: different synthesized videos with diverse gestures

More Results

Ablation for disentangled learning

Method	Diversity ↑	Multimodality ↑
Baseline	8.753	-
+Initial State	8.036	9.943
+Disentangled Training	21.749	23.740
+Random Noise	**23.055**	**24.898**

User study


Lip-sync
Image quality
Diversity
Preference
Speech2Gesture
SDT
Ours