USC Institute for Creative Technologies

ICCV OCT 19-23, 2025 HONOLULU HAWAII

DiTaiListener: Controllable High Fidelity Listener Video Generation with Diffusion

ICCV 2025

Maksim Siniukov*

Di Chang*

Minh Tran

Hongkun Gong

Ashutosh Chaubey

Mohammad Soleymani

Listener Behavior Generation

Generate listener response given speaker's cues: head motions, facial expression, voice

speaker

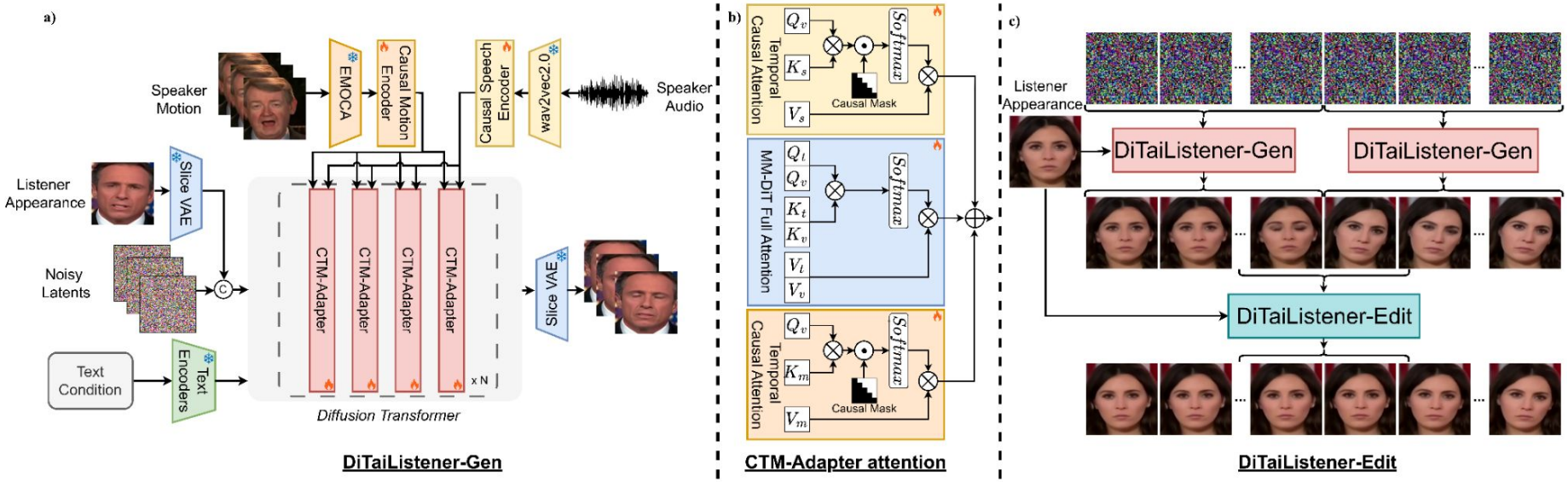
listener

Background & Challenges

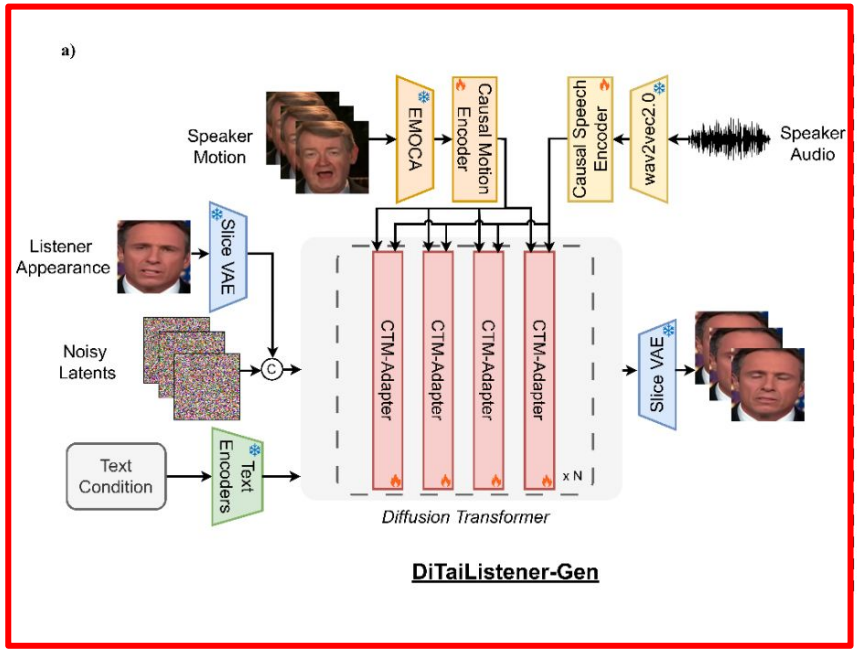
Existing listener generation methods:

- rely on low-dimensional motion codes (3DMMs) with limited expressiveness and visual realism
- work on short clips, causing fragmented transitions and poor long-term coherence
- often lack temporal causality between cues and reactions

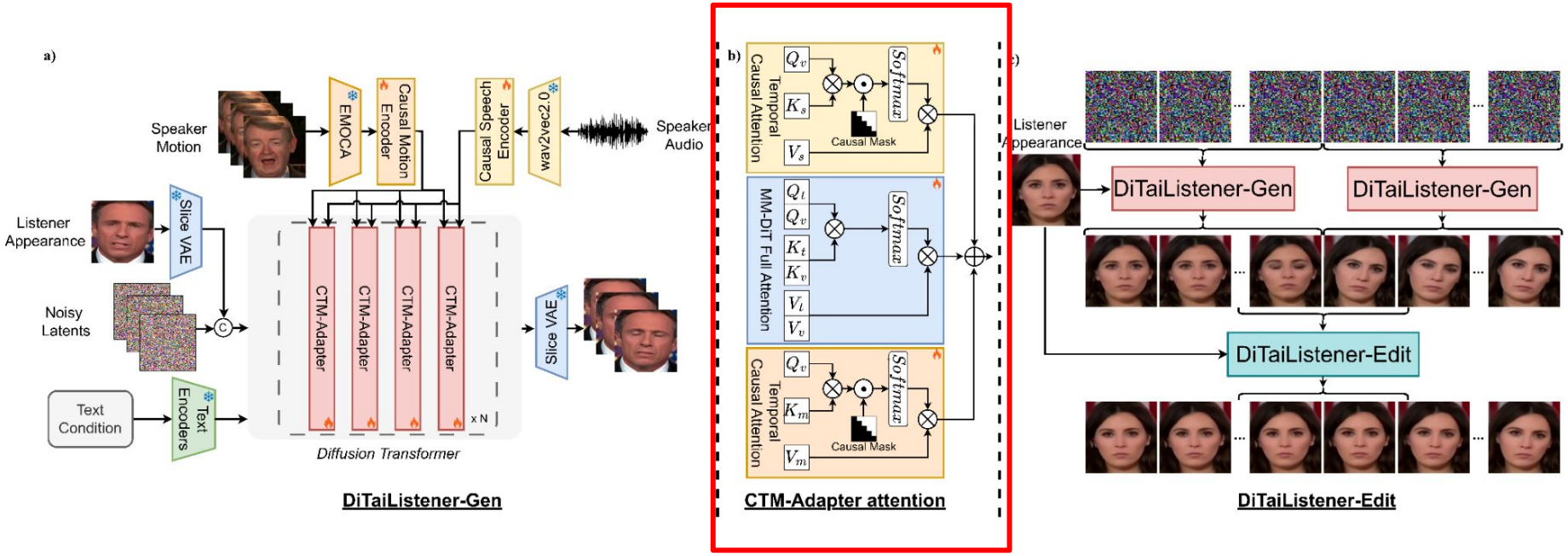
Method


a)
Speaker Motion
EMOCA
Causal Motion Encoder
Causal Speech Encoder
wav2vec2.0
Speaker Audio
Listener Appearance
Slice VAE
Noisy Latents
CTM-Adapter
CTM-Adapter
CTM-Adapter
CTM-Adapter
x N
Slice VAE
Text Condition
Text Encoders
Diffusion Transformer
DiTaiListener-Gen
b)
Temporal Causal Attention
Softmax
Causal Mask
MM-DIT Full Attention
Softmax
Temporal Causal Attention
Softmax
Causal Mask
CTM-Adapter attention
c)
Listener Appearance
DiTaiListener-Gen
DiTaiListener-Gen
DiTaiListener-Edit
DiTaiListener-Edit

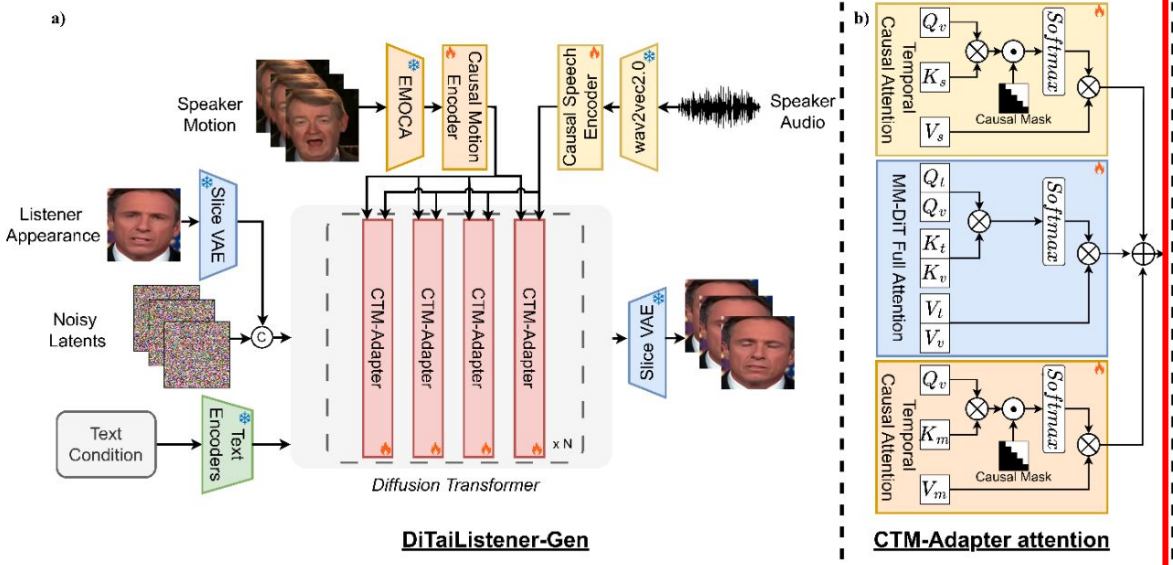
Method


a)
Speaker Motion
EMOCA
Causal Motion Encoder
Causal Speech Encoder
wav2vec2.0
Speaker Audio
Listener Appearance
Slice VAE
Noisy Latents
CTM-Adapter
CTM-Adapter
CTM-Adapter
CTM-Adapter
x N
Slice VAE
Text Condition
Text Encoders
Diffusion Transformer
DiTaiListener-Gen
b)
Temporal Causal Attention
Softmax
Causal Mask
MM-DiT Full Attention
Softmax
Temporal Causal Attention
Softmax
Causal Mask
CTM-Adapter attention
c)
Listener Appearance
DiTaiListener-Gen
DiTaiListener-Gen
DiTaiListener-Edit
DiTaiListener-Edit

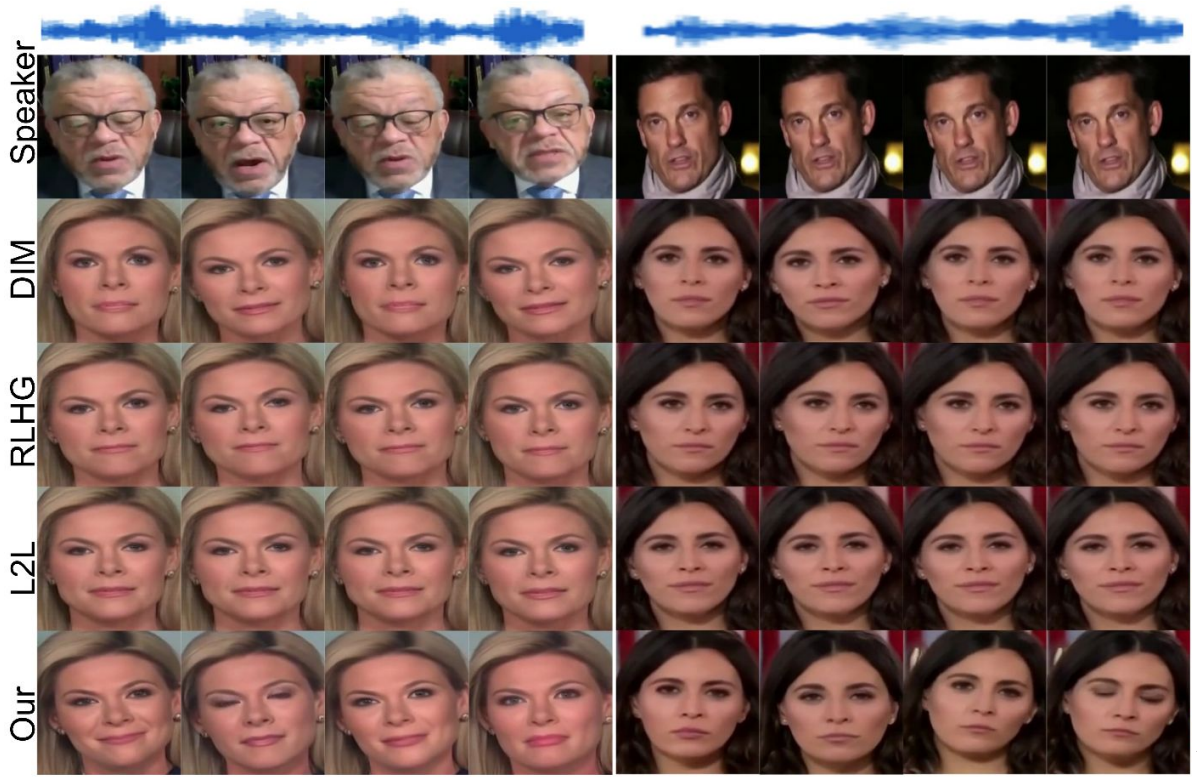
Method


a)
Speaker Motion
EMOCA
Causal Motion Encoder
Causal Speech Encoder
wav2vec2.0
Speaker Audio
Listener Appearance
Slice VAE
Noisy Latents
CTM-Adapter
CTM-Adapter
CTM-Adapter
CTM-Adapter
x N
Slice VAE
Text Condition
Text Encoders
Diffusion Transformer
DiTaiListener-Gen
b)
Temporal Causal Attention
Softmax
Causal Mask
MM-DIT Full Attention
Softmax
Temporal Causal Attention
Softmax
Causal Mask
CTM-Adapter attention
c)
Listener Appearance
DiTaiListener-Gen
DiTaiListener-Gen
DiTaiListener-Edit
DiTaiListener-Edit

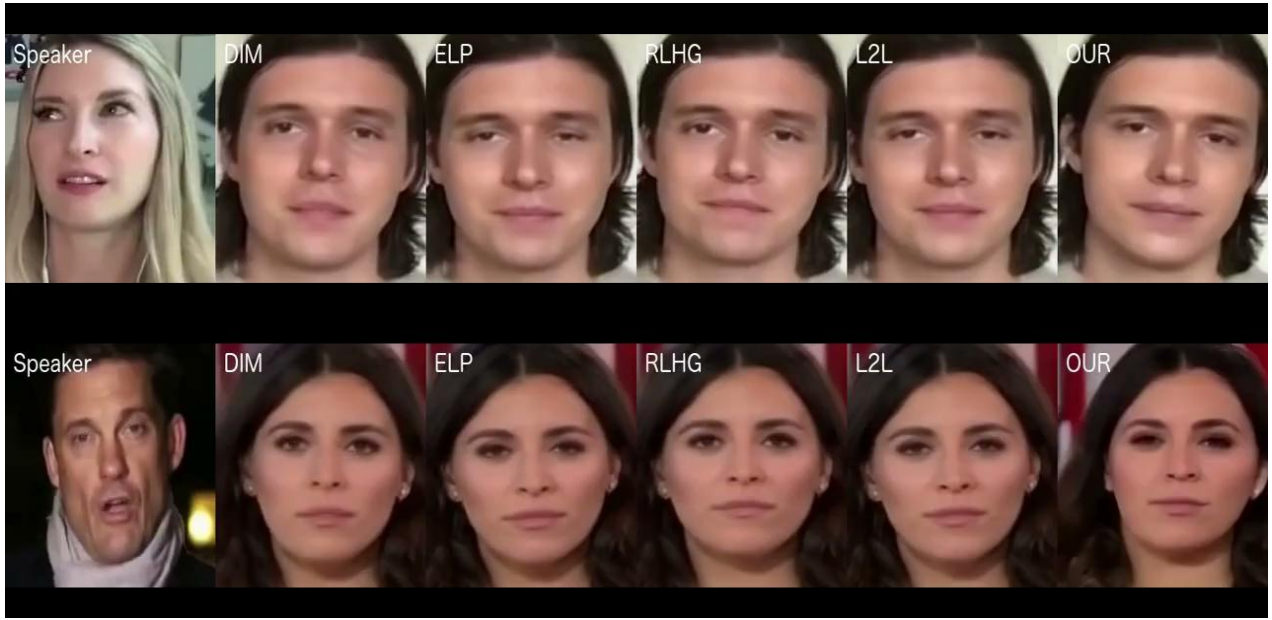
Method


a)
Speaker Motion
EMOCA
Causal Motion Encoder
Causal Speech Encoder
wav2vec2.0
Speaker Audio
Listener Appearance
Slice VAE
Noisy Latents
Text Condition
Text Encoders
CTM-Adapter
CTM-Adapter
CTM-Adapter
CTM-Adapter
x N
Slice VAE
Diffusion Transformer
DiTaiListener-Gen
b)
Temporal Causal Attention
Softmax
Causal Mask
MM-DIT Full Attention
Softmax
Temporal Causal Attention
Softmax
Causal Mask
CTM-Adapter attention
c)
Listener Appearance
DiTaiListener-Gen
DiTaiListener-Gen
DiTaiListener-Edit
DiTaiListener-Edit

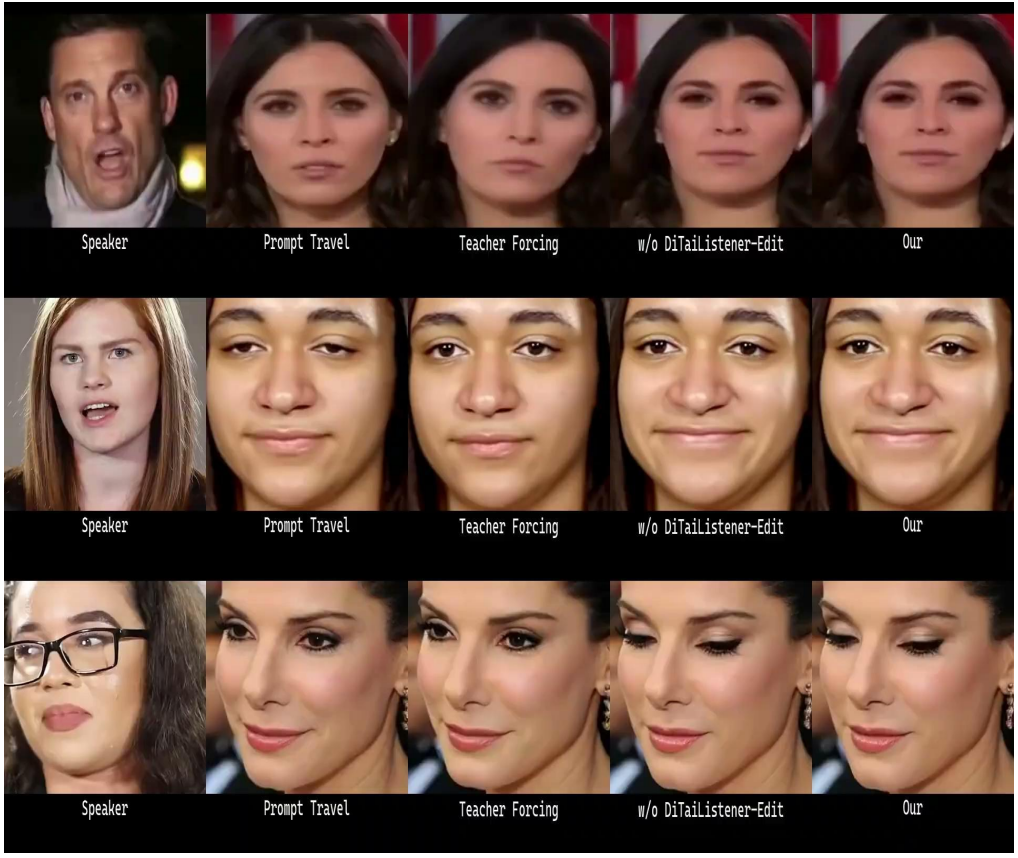
Results. Qualitative Comparison



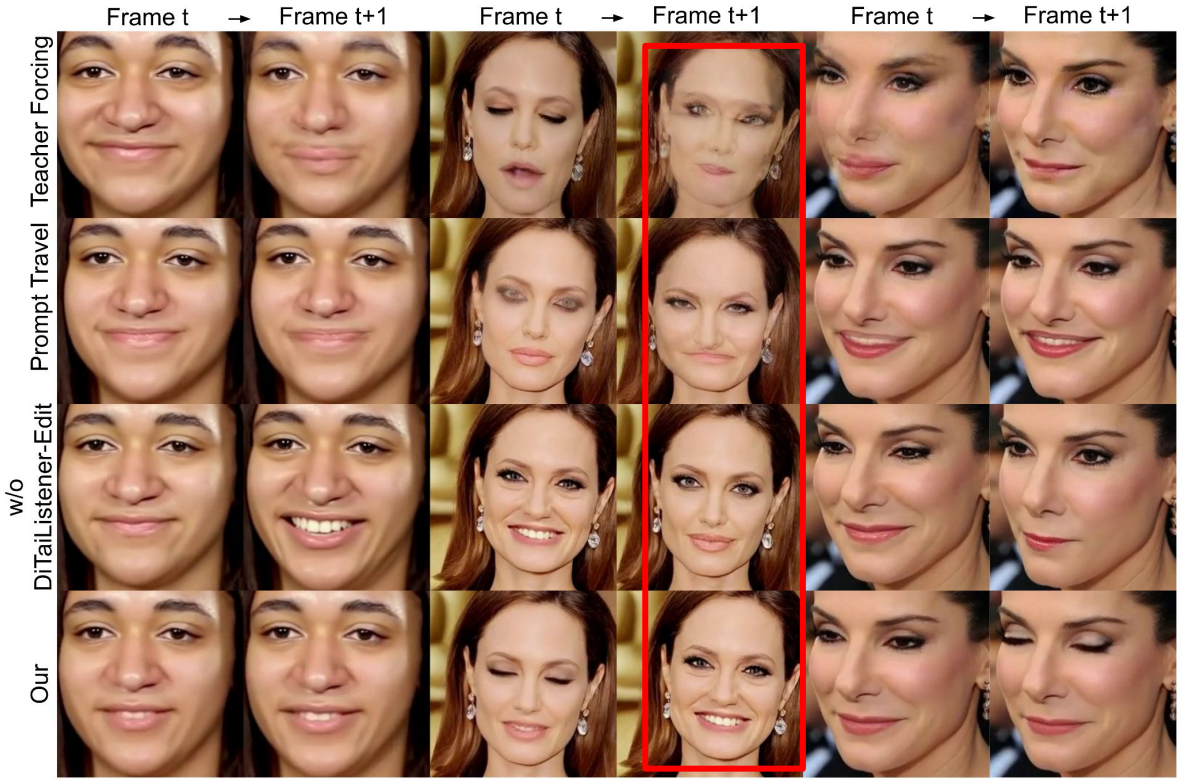
Results. Qualitative Comparison



Results. Long video generation



Results. Long video generation



Results. Quantitative Comparison

RealTalk Dataset

Method	FID↓	FVD↓	LPIPS↓	PSNR↑	SSIM↑
RLHG [71]	30.55	217.48	0.41	17.23	0.53
L2L [39]	56.43	522.67	0.46	16.49	0.53
DIM [56]	30.75	204.49	0.40	17.74	0.54
w/o Text-Control	8.05	49.88	0.27	18.70	0.59
w/o CTM-Adapter	8.16	45.62	0.27	18.60	0.55
DiTaiListener	**7.99**	**45.42**	**0.26**	**18.73**	**0.60**

USC Institute for Creative Technologies

Quantitative Comparison in 3DMM-space

VICO Dataset

Method	FD↓		P-FD↓		MSE↓		SID↑		Var↑	
	Exp	**Pose**	**Exp**	**Pose**	**Exp**	**Pose**	**Exp**	**Pose**	**Exp**	**Pose**
RLHG [81]	69.04	0.05	69.09	0.06	1.37	**0.01**	0.35	3.23	0.14	**0.01**
L2L [40]	72.89	0.10	72.94	0.10	1.44	0.02	0.10	2.42	0.07	**0.01**
DIM [56]	77.97	0.15	78.70	0.15	1.52	0.02	3.49	3.29	0.74	**0.01**
GT	-	-	-	-	-	-	5.13	3.95	1.36	0.02
w/o Text-Control	15.62	**0.02**	16.35	**0.02**	0.67	**0.01**	5.01	3.94	1.30	**0.01**
w/o CTM-Adapter	15.10	**0.02**	15.82	**0.02**	0.66	**0.01**	**5.11**	**3.98**	1.30	**0.01**
DiTaiListener	**14.28**	**0.02**	**15.07**	**0.02**	**0.65**	**0.01**	5.09	3.95	**1.31**	**0.01**

Results. User Study

User study. We ask the participants to choose the best video among all methods

Method	Feedback↑	Diversity↑	Smoothness↑	Overall↑
ELP [49]	2.93%	4.04%	5.65%	4.21%
RLHG [81]	10.45%	12.14%	13.00%	11.86%
L2L [39]	4.20%	6.06%	10.95%	7.07%
DIM [56]	10.25%	12.12%	16.11%	12.83%
Ours	**72.17%**	**65.64%**	**54.29%**	**64.03%**

Results. Out-of-domain generation

Results. Text-Conditioned Listener Generation

Summary

- Introduced pixel-space **video diffusion** model for listener response generation, achieving **73.8%** improvement in photorealism
- Proposed DiTaiListener-Edit: coherent **long video** generation method
- Introduced Causal Temporal **Multimodal Adapter** (CTMAdapter) for diffusion transformer that allows controllable face video generation and **aligns** listener responses with speaker cues
- Demonstrated the potential of video generation methods for socially intelligent, coordinated, and controlled face and head gestures
- Allowed free-form text control to guide the listener's behaviors

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant IIS-2211550. The work was also sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF20-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office, or the U.S. Government and do not necessarily reflect the views of the National Science Foundation. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

USC Institute for Creative Technologies