DGTalker: Disentangled Generative Latent Space Learning for Audio-Driven Gaussian Talking Heads

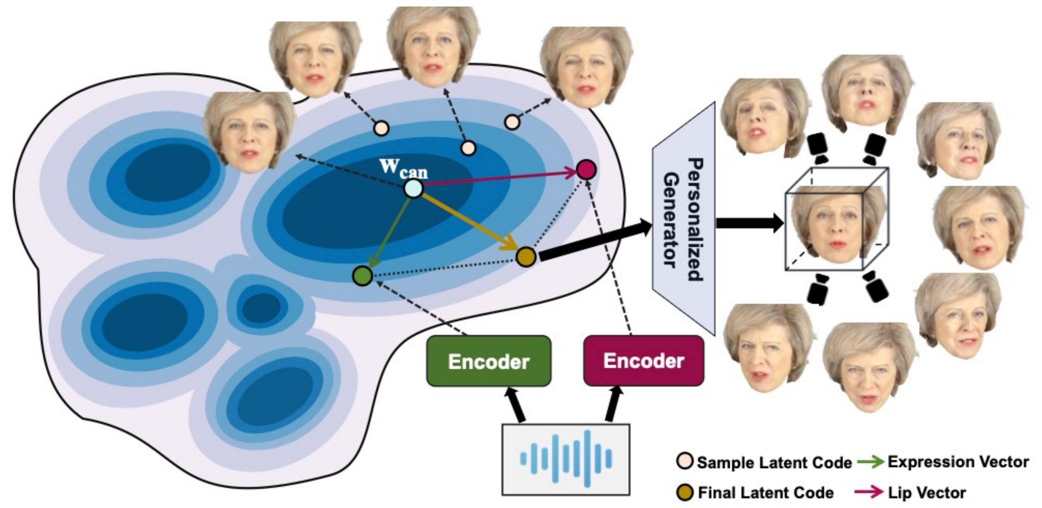
Xiaoxi Liang[1], Yanbo Fan[2], Qiya Yang[1], Xuan Wang[3], Wei Gao[1], Ge Li[1]
[1]Peking University [2]Nanjing University [3]Ant Group

ICCV OCT 19-23, 2025 HONOLULU HAWAII

INTRODUCTION

What We Do

Achieve **real-time**, **high-fidelity**, and **broader rendering perspective** talking head synthesis from monocular videos.



Motivation

A highly practical talking 3D head avatar needs to meet the following technical requirements:

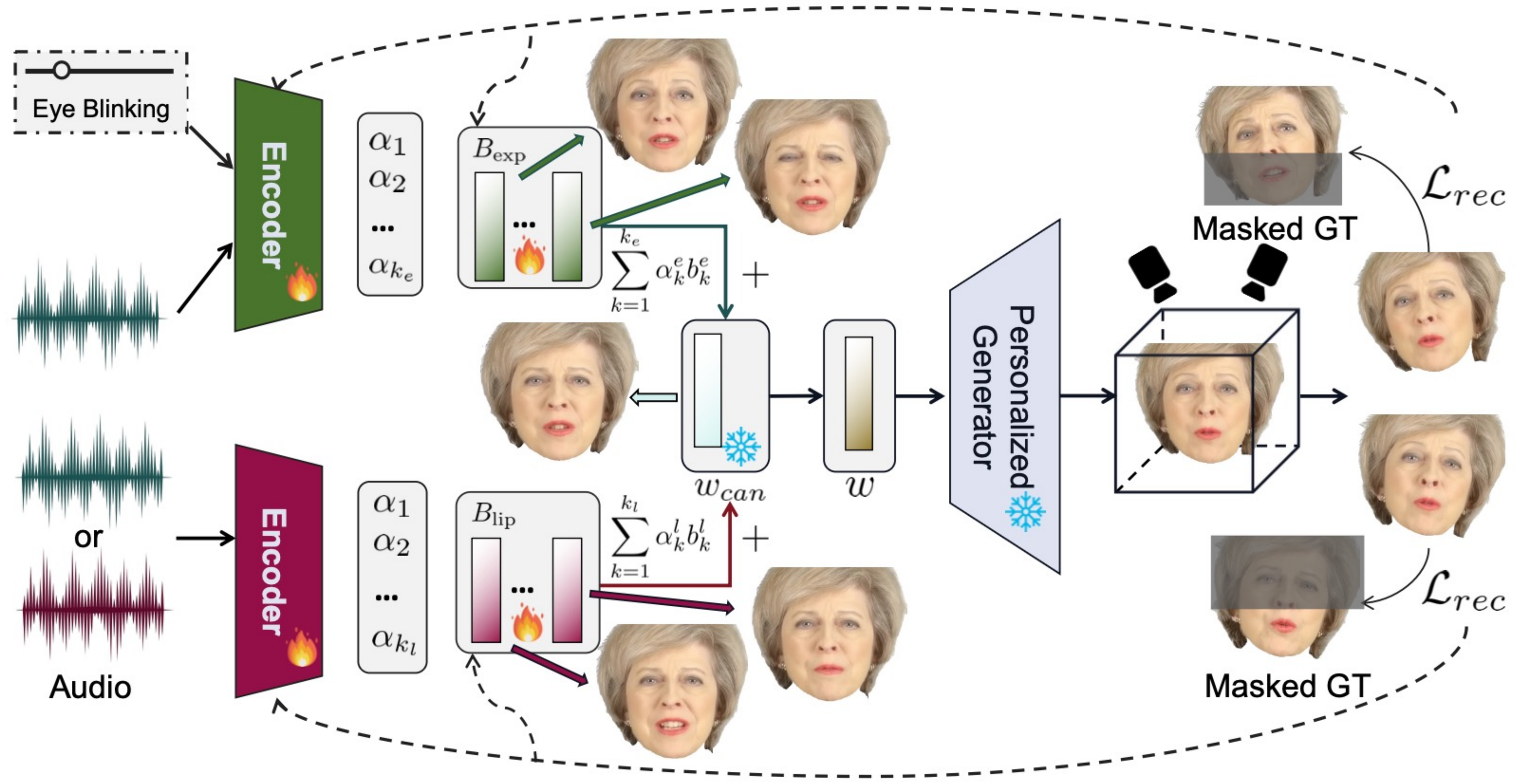
- Building from **monocular talking videos**, which are easier to obtain from consumers
- High **real-time** inference performance, high-fidelity **rendering quality**
- Visual quality from **broader viewpoints**

Early methods suffer from poor geometry, appearance quality, and inadequate rendering robustness across views, due to direct adoption of Vanilla 3DGS in monocular scenarios.

Contribution

- Leverage generative priors and formulate the task as latent space navigation
- Propose a **disentangled framework** for audio-generator modality mismatch
- Introduce a **masked cross-view supervision** strategy to ensure disentangled learning

METHOD



Our goal is to find **the optimal latent code $\boldsymbol{\omega}$** in the generator's space, conditioned on the given audio. We decompose $\boldsymbol{\omega}$ into:

- ω_{can}: encodes a **global** canonical expression for a **specific identity**
- two sets of learnable blendshapes B_{exp}, B_{lip}: characterize the **expression variations of the upper and lower face**, respectively

Dual-audio encoders are employed to regress the blendshapes coefficients.

In training stage, we sometimes generate a non-existent head by combing the lip code from one audio and the expression code from another, and render the 3DGS head under each **audio-correlated viewpoint**, and apply **region-specific supervision** focusing on the upper/lower face, respectively.

Visualization

Expression Blendshapes

Lip Blendshapes

Controllability

EXPERIMENTS

Ablation Studies

Method	PSNR ↑	LPIPS ↓	FID ↓	LMD ↓	Sync ↑
Ground Truth	N/A	0	0	0	8.468
w/o Disentangled Design	27.741	0.101	19.951	4.547	3.869
w/o Dual-Encoders	28.473	0.073	16.208	4.127	5.870
w/o Blendshapes	28.868	0.070	15.156	4.191	6.189
w/o MCS	28.559	0.072	15.742	4.551	4.547
All	**28.943**	**0.065**	**15.149**	**3.997**	**6.295**

Comparison to SOTA

Methods	FID ↓	IDSIM ↑	AUE ↓	Sync-E ↓	Sync-C ↑
Ground Truth	0	1	0	6.859	8.468
ER-NeRF	228.740	0.306	2.753	11.141	2.887
GaussianTalker	138.332	0.337	<u>2.601</u>	10.785	3.773
TalkingGaussian	137.914	0.363	2.621	<u>9.624</u>	<u>5.198</u>
HFA-GP*	<u>99.601</u>	<u>0.373</u>	2.745	12.455	1.627
Ours	**80.011**	**0.436**	**2.525**	**9.565**	**5.255**

Novel-view self-driven

Method	Test Audio A		Test Audio B	
	Sync-C ↑	Sync-E ↓	Sync-C ↑	Sync-E ↓
Ground Truth	8.167	6.808	8.080	7.182
ER-NeRF	2.267	11.669	2.458	11.369
GaussianTalker	3.242	10.903	1.557	12.476
TalkingGaussian	<u>3.922</u>	<u>10.528</u>	<u>3.999</u>	<u>10.202</u>
HFA-GP*	1.098	13.172	0.976	12.966
Ours	**3.947**	**10.433**	**4.069**	**10.106**

Generalize to unseen audio

Self-driven **Novel-view self-driven** **Unseen audio clip**

CONCLUSION

- We introduce **DGTalker**, a novel framework for real-time, high-fidelity audio-driven Gaussian talking head synthesis.
- **DGTalker** achieves SOTA performance with extra controllability.