SwapTalk: Audio-Driven Talking Face Generation with One-Shot Customization in Latent Space

Zeren Zhang*, Haibo Qin*, Jiayu Huang, Jo-Ku Cheng, Yixin Li, Hui Lin, Yitao Duan, Jinwen Ma
Institution: Peking University, Youdao AI
* Equal contribution

ICASSP 2025 Best Student Paper Award

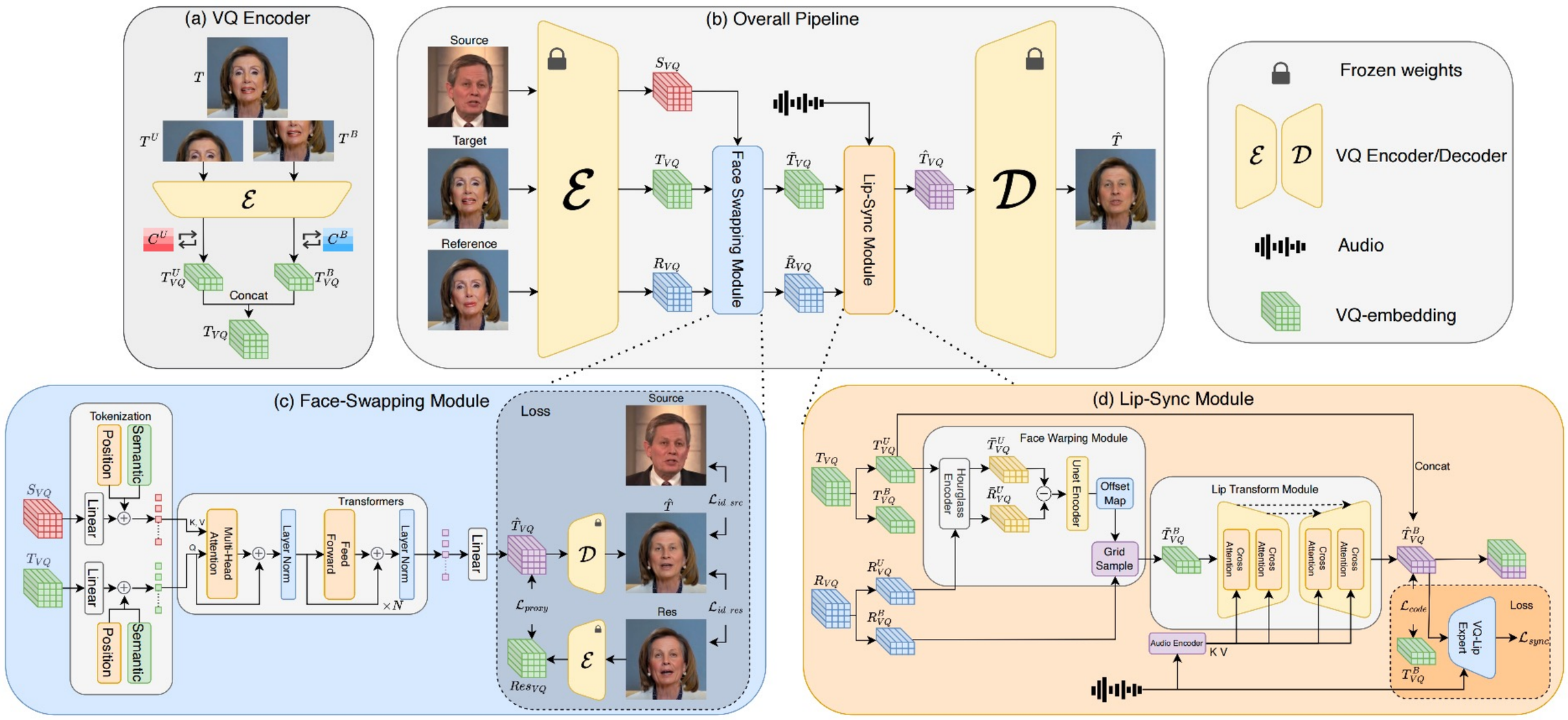
INTRODUCTION ▸ ▸ ▸

- **Background:** Combining face swapping with lip synchronization provides a cost-efficient solution for generating customized digital avatars.
- **Problem:** Existing models struggle with face swapping and lip synchronization interference when cascaded in RGB space.
- **Solution:** SwapTalk, a unified framework operating in the VQ-embedding latent space, effectively addresses these issues while maintaining high fidelity and synchronization.

Contribution

- We propose a unified framework that completes face-swapping and lip-sync tasks within a semantically rich and decoupled VQ-embedding space, simultaneously achieving lip-sync and preserving both ID appearance and overall consistency.
- We utilize the identity loss during the training process of the pre-trained VQGAN and the face-swapping module to enhance the generalization of unseen identities and consistency across time series. Additionally, we employ a expert supervision within the VQ-embedding space to improve lip-sync accuracy.
- We point out issues with lip-sync evaluation in prior research and propose the use of an audio-visual un-synchronized setting to more accurately assess performance in realistic scenarios. Furthermore, we introduce a metric designed to reasonably evaluate facial identity consistency within videos.

APPROACH ▸ ▸ ▸



EXPERIMENTAL RESULTS ▸ ▸ ▸

Task Type	Methods	FID↓	SSIM↑	CPBD↑	LMD↓	LSE-C↑	ID Retrieve↑	Consistency↑
Self-Driven	Ground Truth	-	-	0.536	-	8.37	-	-
	Wav2Lip[127]	28.5	0.813	0.425	1.959	**9.50**	-	-
	Wav2Lip-Restore	15.1	0.904	0.535	1.624	9.04	-	-
	Sync-Swap	12.4	0.904	0.484	1.481	7.49	82.9	76.00
	Sync-Swap-Restore	12.7	0.885	**0.539**	1.483	7.25	81.4	75.57
	Swap-Sync	11.7	0.906	0.482	1.384	8.89	80.3	74.89
	Swap-Restore-Sync	12.5	0.890	0.533	1.386	8.84	79.3	74.38
	WAVSYNCSWAP[131]	49.9	0.738†	0.470	3.161	9.09	85.7	64.17†
	SwapTalk	11.6	0.908	0.521	1.221	9.08	87.8	78.00
	SwapTalk (Extra Data)	**11.1**	**0.910**	0.530	**1.139**	9.25	**92.3**	**81.88**
Cross-Driven	Wav2Lip[127]	27.4	0.811	0.417	-	7.87	-	-
	Wav2Lip-Restore	14.6	0.877	**0.533**	-	7.63	-	-
	Sync-Swap	12.7	0.899	0.483	-	7.52	84.2	77.83
	Sync-Swap-Restore	13.0	0.885	0.528	-	7.24	83.8	77.44
	Swap-Sync	11.6	0.900	0.484	-	8.62	82.6	76.63
	Swap-Restore-Sync	11.8	0.888	0.521	-	8.57	80.4	75.99
	SwapTalk	11.0	0.905	0.524	-	8.94	87.6	80.19
	SwapTalk (Extra Data)	**10.8**	**0.907**	0.526	-	**8.99**	**93.5**	**82.57**

TABLE V
THE IMPACT OF DIFFERENT BACKBONES.

Lip-Sync Expert	Backbone	FID↓	SSIM↑	LMD↓
	UNet from [39]	4.3	0.940	1.381
✓	UNet from [39]	**4.2**	**0.942**	**1.009**
✓	UNet	5.5	0.938	1.252
✓	DiT	6.1	0.936	1.160

TABLE II
THE IMPACT OF VQGAN WITH DIFFERENT SPATIAL COMPRESSION RATIOS ON FACE SWAPPING PERFORMANCE IN HDTF DATASET.

Spatial Compress Rate	ID Retrieve↑	FID↓
8×	85.63	15.6
16×	**94.29**	**9.5**

TABLE III
THE IMPACT OF VQGAN WITH DIFFERENT SPATIAL COMPRESSION RATIOS ON LIP SYNCHRONIZATION AND VIDEO QUALITY IN HDTF DATASET.

Spatial Compress Rate	FID↓	SSIM↑	LMD↓
8×	8.3	0.841	1.116
16×	**4.2**	**0.942**	**1.008**

TABLE IV
PERFORMANCE OF DIFFERENT VARIANTS OF FACE SWAPPING MODULES.

$\mathcal{L}_{id\_src}$	$\mathcal{L}_{id\_res}$	Backbone	ID Retrieve↑	FID↓
		Transformer	72.3	16.0
✓		Transformer	93.6	15.5
✓	✓	Transformer	**94.3**	**9.5**
✓	✓	UNet	80.1	15.7
✓	✓	UNet from [6]	84.6	12.3