JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

LipFormer: High-fidelity and Generalizable Talking Face Generation with A Pre-learned Facial Codebook

Jiayu Wang[1] Kang Zhao[1] Shiwei Zhang[1] Yingya Zhang[1] Yujun Shen[2] Deli Zhao[1] Jingren Zhou[1]

[1]Alibaba Group [2]Ant Group

{wangjiayu.wjy, zhaokang.zk, zhangjin.zsw, yingya.zyy, jingren.zhou}@alibaba-inc.com,
{shenyujun0302, zhaodeli}@gmail.com

LipFormer

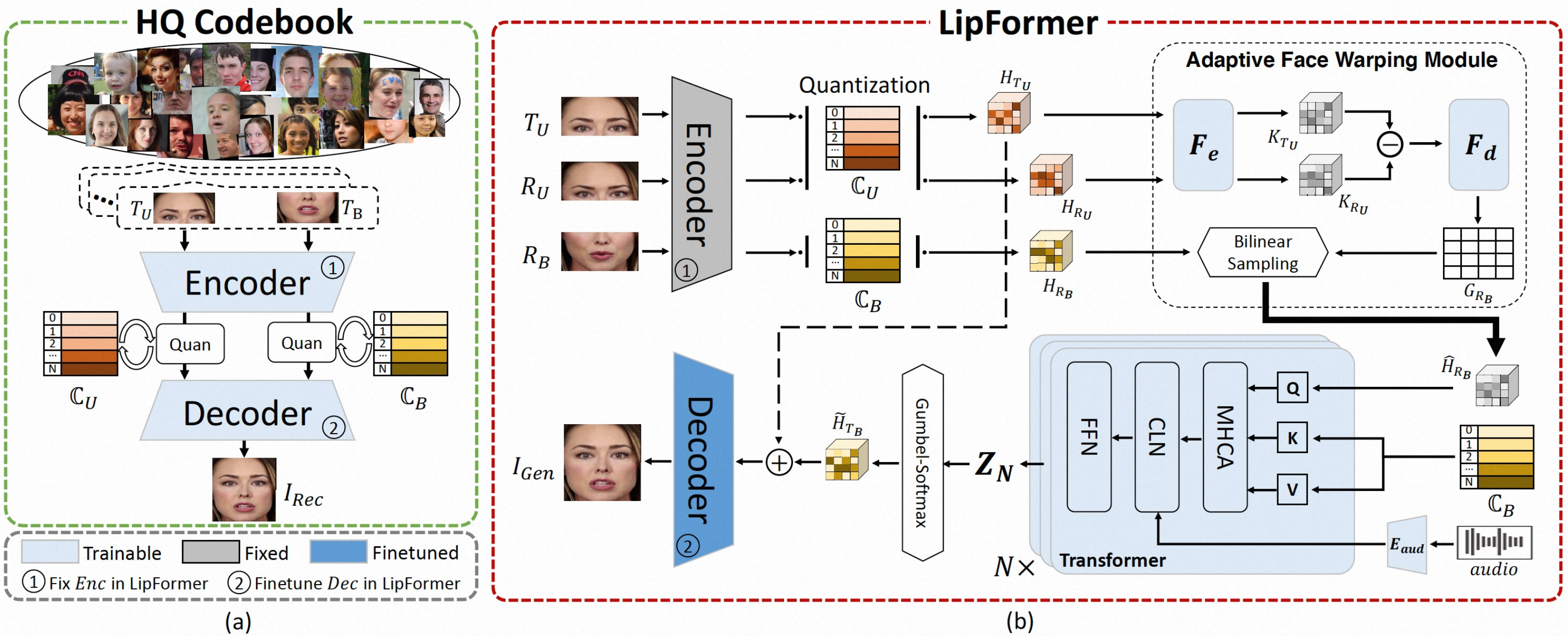


Figure 2. Overview of the proposed LipFormer. (a) HQ Codebook Learning (Sec. 3.1). A quantized autoencoder is trained with face reconstruction task, which outputs two codebooks. (b) LipFormer Training (Sec. 3.2). We fix the face encoder and the codebooks, and finetune the decoder with other parts end to end. Conditioned on the input audio and a reference face, the Transformer module is introduced to predict the target lip-codes. Moreover, an adaptive face warping module is designed to address the texture mismatch issue.

(a) Representing diverse face details

(b) Finding proper lip-codes

Experiments Results

Talking Face Synthesis

- Methods:

  1. Reconstruction based methods (Wav2Lip)

  2. Implicit representation methods (AD-NeRF)

- Limitations:

  1. Low resolution and qualities (LRW and LRS2), leading to the learned model an unsatisfying synthesis quality

  2. A limited number of identities (Obama), which requires training a specific model for each person and it is hard to generalize to unseen portraits

Talking Face Synthesis

There are many publicly available datasets of high-resolution face images, e.g., the FFHQ dataset contains 70,000 identities with 1024×1024 resolutions

Could these image datasets benefit the generation of a talking portrait ?

✓ The answer is a big yes!

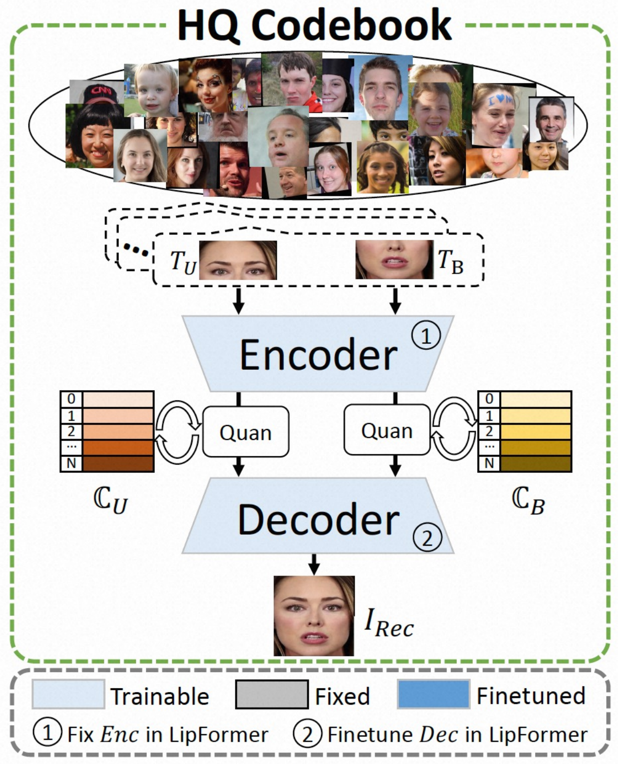
LipFormer

- HQ Codebook

  ➢ This stage aims to learn the codebooks, so that they can be retrieved to generate HQ talking face images

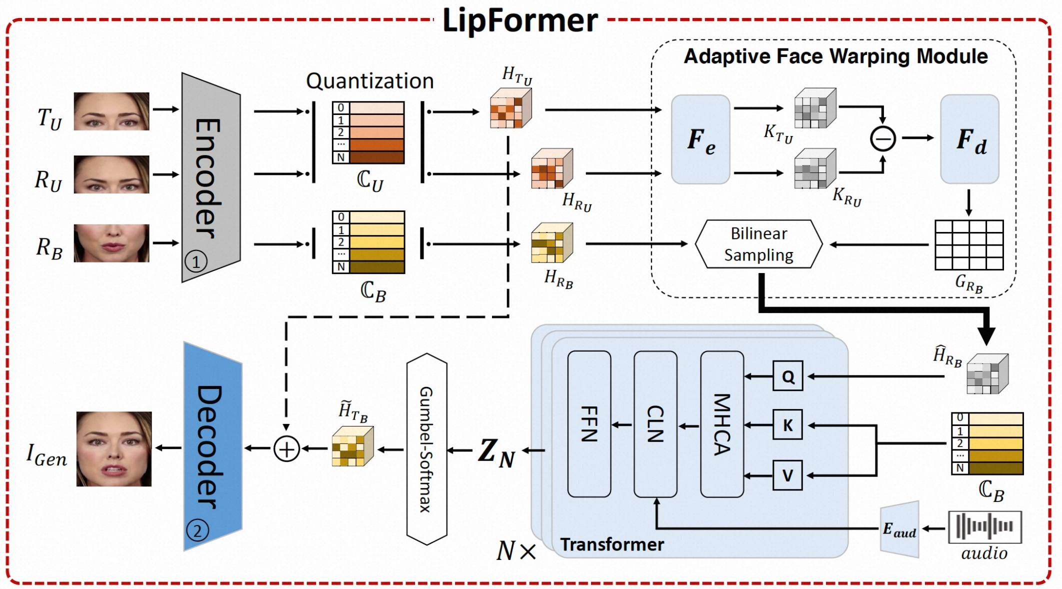
$$\mathcal{L}_{\text{VQ}} = \|\text{sg}[Enc(\boldsymbol{T}_U)] - \boldsymbol{H}_U\|_2^2 + \beta \|\text{sg}[\boldsymbol{H}_U] - Enc(\boldsymbol{T}_U))\|_2^2 + \|\text{sg}[Enc(\boldsymbol{T}_B)] - \boldsymbol{H}_B\|_2^2 + \beta \|\text{sg}[\boldsymbol{H}_B] - Enc(\boldsymbol{T}_B))\|_2^2 ,$$

$$\mathcal{L}_{Rec} = \mathcal{L}_{\text{VQ}} + \mathcal{L}_2^{Rec} + 0.1\mathcal{L}_{per}^{Rec} + 0.1\mathcal{L}_{adv}^{Rec} .$$



(a)

LipFormer



(b)

- LipFormer
  - ➢ This stage aims to predict lip-codes

$$\mathcal{L}_{Gen} = \lambda_{Tr}\mathcal{L}_{Tr} + \mathcal{L}_2^{Gen} + \lambda_{per}\mathcal{L}_{per}^{Gen} + \lambda_{adv}\mathcal{L}_{adv}^{Gen},$$

Experiments

Table 1. The quantitative results on LRS2 and our collected YouTubeHQ. We compare the proposed LipFormer against several baseline methods. We adopt PSNR and SSIM to measure image quality, LMD to measure mouth shape coherence, LSE-D and LSE-C to measure lip-sync quality.

	LRS2					YouTubeHQ		
Methods	PSNR(↑)	SSIM(↑)	LMD(↓)	LSE-D(↓)	LSE-C(↑)	PSNR(↑)	SSIM(↑)	LMD(↓)
Ground Truth	N/A	1.000	0.000	6.259	8.247	N/A	1.000	0.000
ATVG [18]	30.427	0.735	2.549	8.223	5.584	24.036	0.707	3.146
Wav2Lip [25]	31.274	0.837	1.940	**5.995**	**8.797**	25.971	0.758	2.473
PC-AVS [51]	29.887	0.747	1.963	7.301	6.728	25.106	0.714	2.606
SyncTalkFace [24]	32.529	0.876	1.387	6.352	7.925	-	-	-
LipFormer	**33.497**	**0.891**	**1.261**	6.408	7.874	**33.249**	**0.876**	**1.357**

		YouTubeHQ/ LRS2				LRW/ LRS3/ HDTF						
	FPS↑	LSE-D↓	LSE-C↑	FID↓	CPBD↑	PSNR↑	SSIM↑	LMD↓	LSE-D↓	LSE-C↑	FID↓	CPBD↑
ATVG	**36.13**	9.65	4.03	12.87/ 8.04	0.22/ 0.20	31.09/ 27.87/ 24.86	0.77/ 0.71/ 0.71	2.03/ 3.14/ 3.14	7.87/ 9.04/ 9.58	5.71/ 4.40/ 4.22	6.41/ 9.34/ 12.63	0.12/ 0.18/ 0.19
Wav2Lip	32.05	**7.68**	**5.57**	11.15/ 4.78	0.23/ 0.27	32.27/ 30.11/ 26.37	0.87/ 0.83/ 0.77	1.41/ 1.98/ 2.26	**6.62**/ **6.67**/ 7.90	**7.15**/ **8.90**/ **5.23**	2.74/ 4.53/ 10.04	0.15/ 0.27/ 0.21
PC-AVS	4.63	8.31	5.28	12.33/ 9.22	0.21/ 0.21	29.39/ 27.84/ 25.22	0.76/ 0.72/ 0.72	1.61/ 2.99/ 2.51	7.55/ 8.16/ 8.19	6.20/ 5.81/ 4.83	7.04/ 9.83/ 12.82	0.10/ 0.19/ 0.20
LipFormer	9.92	7.71	5.48	**3.93**/ **3.76**	**0.29**/ **0.29**	**33.83**/ **32.93**/ **33.26**	**0.90**/ **0.87**/ **0.87**	**1.26**/ **1.38**/ **1.34**	6.96/ 6.89/ **7.89**	6.71/ 8.10/ 5.17	**2.38**/ **3.79**/ **3.85**	**0.18**/ **0.28**/ **0.29**

Table B. We add 1) LRW,LRS3,HDTF, 2) missing metrics for LRS2,YouTubeHQ, 3) FPS. SyncTalkFace is ignored for code unavailable.

Experiments

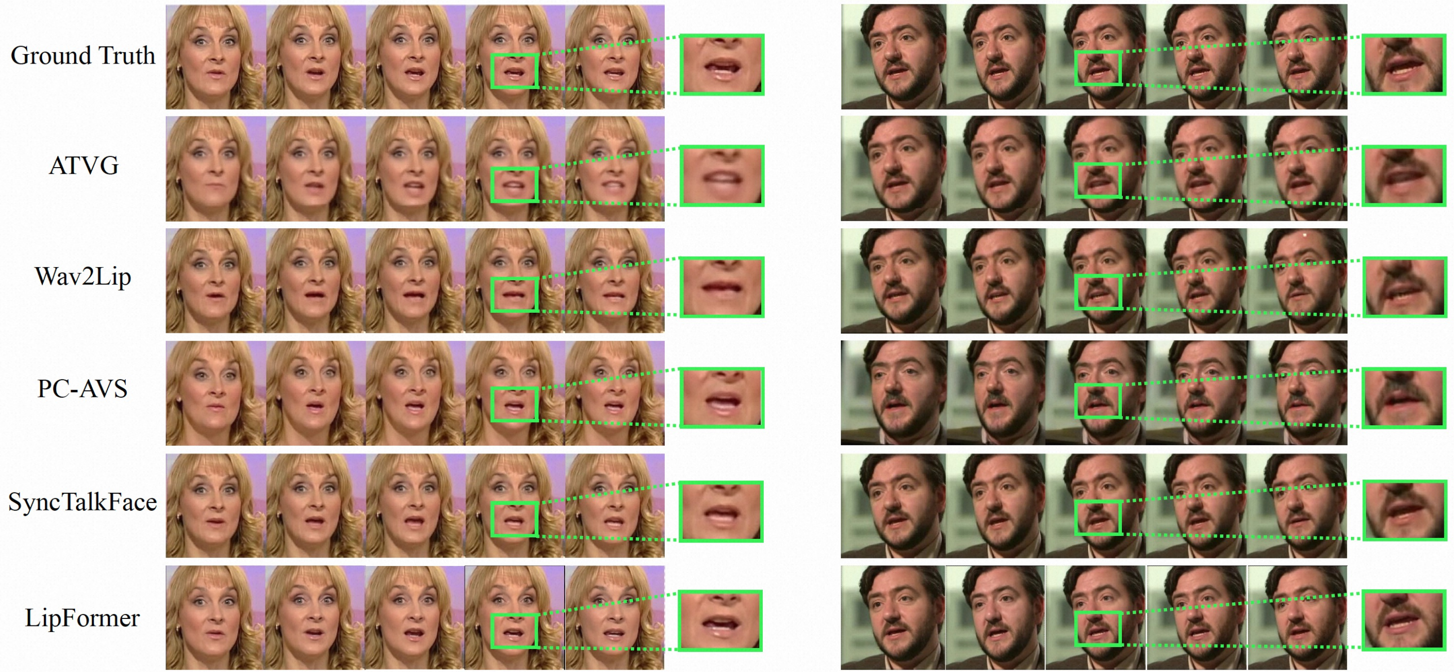


Figure 4. Comparison with other baseline methods for talking face generation on LRS2. Our method generates results that best match the ground truth, and with clear details especially in the mouth region.

Experiments

Table 2. The quantitative results on video samples provided by AD-NeRF [10]. We compare the proposed LipFormer to AD-NeRF. The best result in each metric is highlighted in bold.

	AD-NeRF Video Sample				
Methods	PSNR↑	SSIM↑	LMD↓	LSE-D↓	LSE-C↑
AD-NeRF [10]	29.714	0.842	1.506	6.603	7.542
LipFormer	**33.145**	**0.870**	**1.359**	**6.377**	**7.902**

Figure 5. The comparison of generated frame results on AD-NeRF [10] sample video. Results of AD-NeRF [10], SSP-NeRF [19] and our proposed LipFormer are provided. Our method generates results with higher fidelity and more accurate mouth shape.

Experiments

Models	LRS2		YouTubeHQ	
	PSNR(↑)	SSIM(↑)	PSNR(↑)	SSIM(↑)
Baseline Model	31.613	0.843	28.035	0.749
+ FFHQ Pre-training	32.630	0.873	31.980	0.845
+ Adaptive Warping	32.411	0.865	31.637	0.833
+ FFHQ pre-training & Adaptive Warping	33.497	0.891	33.249	0.876

Table A. Ablation study of FFHQ Pre-training and the Adaptive Face Warping Module.

Variants	LSE-D↓	LSE-C↑
w/o AW	7.91	5.36
w/o FFHQ pt	8.15	5.24
LipFormer	7.71	5.48

Table C. Lip-sync metrics.

n	PSNR↑	SSIM↑	LSE-D↓	LSE-C↑
2048	32.86	0.86	7.76	5.40
4096	**33.25**	**0.88**	**7.71**	**5.48**
8192	31.98	0.84	7.94	5.29

Table D. Ablation of codebook size.

Metrics	Wav2Lip	LipFormer
lip-sync↑	**3.24**	2.74
lip-quality↑	1.12	**2.97**
lip-artifacts↓	3.88	**2.09**

Table E. User Study.

Experiments

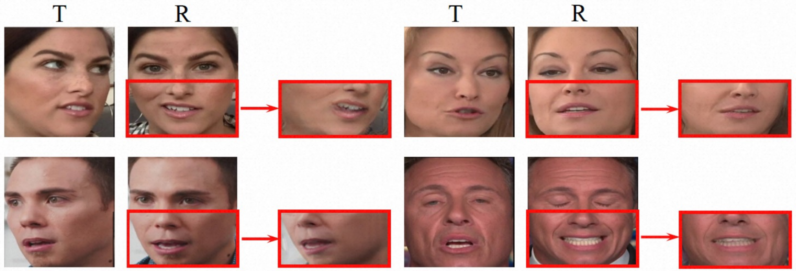


Figure 7. Visualizing warped lip features by directly sending them into the decoder. These visualizations reflect that our proposed face-warping module is effective in facial texture aligning.

Figure B

Experiments

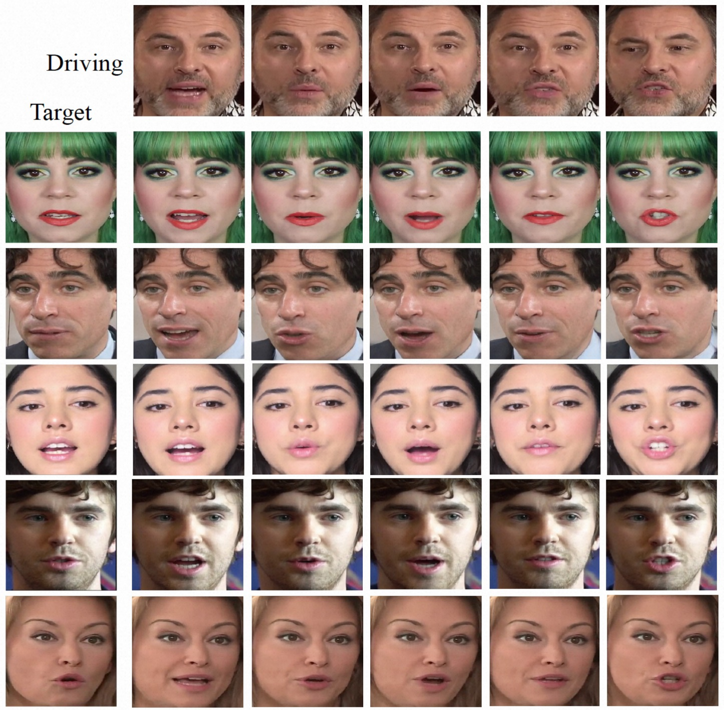


Figure 6. Visual results of mouth shape transferring experiment on our collected YouTubeHQ. The audio feature of each driving video frame is taken to drive each target frame. Each generated result has a mouth shape corresponding to the driving audio.

Experiments

Figure 1. High-fidelity talking face generation with LipFormer. **Top:** Five target face pairs. **Middle:** LipFormer-generated results, driven by target face's own audio. **Bottom:** LipFormer-generated results, after exchanging the audio of each target pair. It is clear that LipFormer successfully captures the relationship between voice and mouth shape.

Q&A