NUS National University of Singapore | School of Computing

CS5647: Sound and Music Computing

Singing Voice Synthesis with Avatar Generation

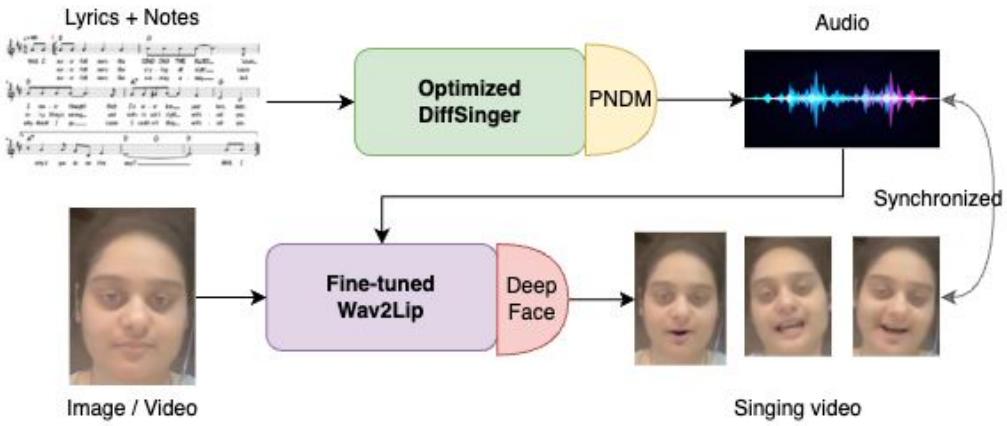
Team 6

Singing Face Generation

1. **Voice Synthesis**
   1. Baseline DiffSinger
   2. Better Vocoder (BigVGAN)
   3. Faster Inference using PNDM

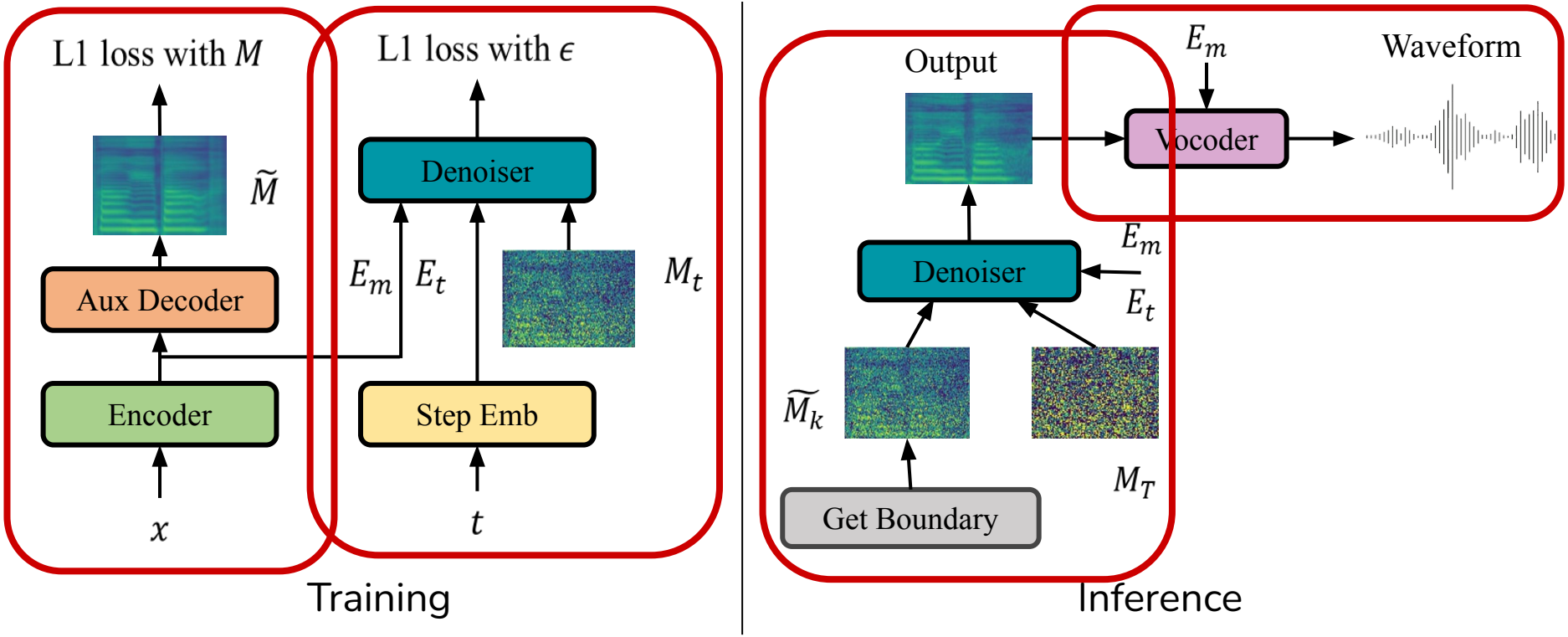
2. **Face Synthesis**
   1. Baseline Wave2Lip
   2. Fine-tune for singing
   3. DeepFace augmentation

3. **Dataset**
   1. OpenCPop
   2. URSing (Audio-Visual Solo Singing)
   3. Own dataset (me singing!)

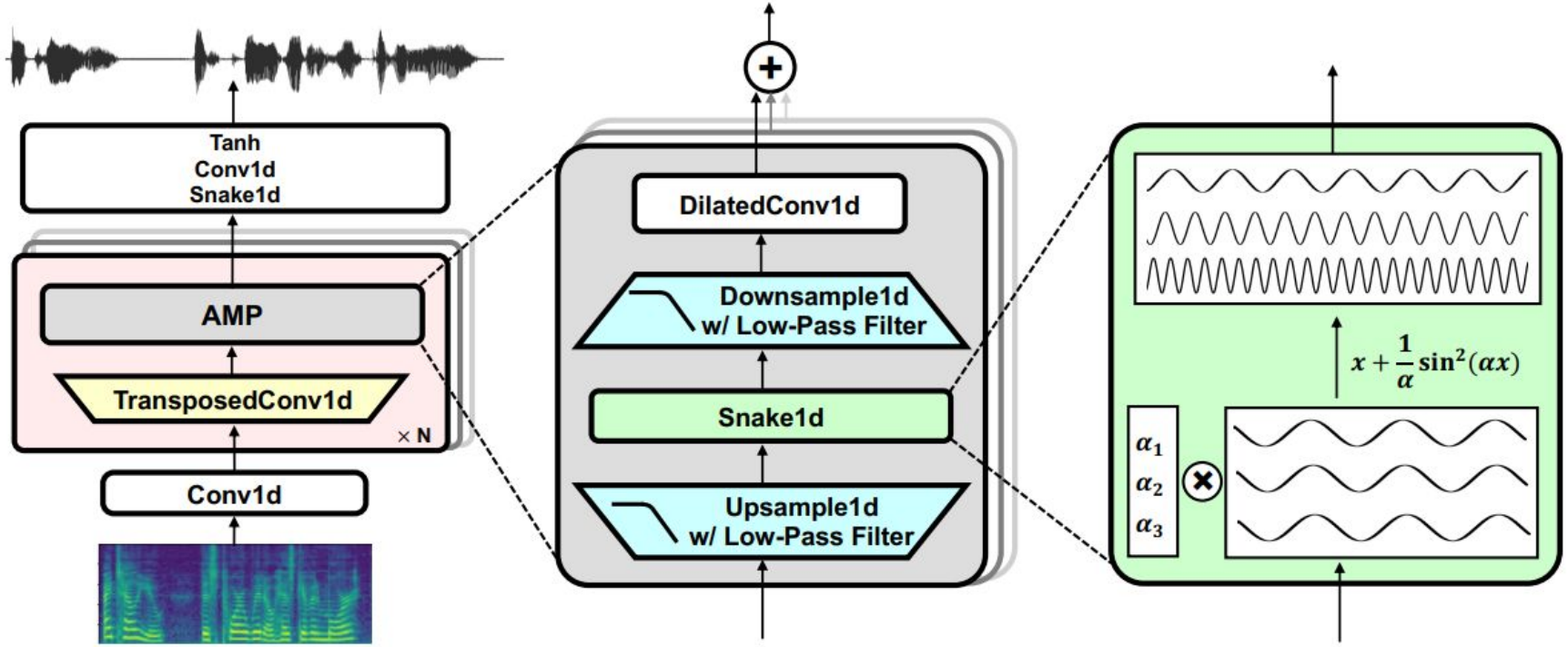


Voice Synthesis

DiffSinger


L1 loss with M
$\widetilde{M}$
Aux Decoder
Encoder
x
L1 loss with ϵ
Denoiser
E_m
E_t
M_t
Step Emb
t
Training
Output
E_m
Waveform
Vocoder
Denoiser
E_m
E_t
$\widetilde{M}_k$
M_T
Get Boundary
Inference

BigVGAN


Tanh
Conv1d
Snake1d
AMP
TransposedConv1d
× N
Conv1d
DilatedConv1d
Downsample1d
w/ Low-Pass Filter
Snake1d
Upsample1d
w/ Low-Pass Filter
$x + \frac{1}{\alpha}\sin^2(\alpha x)$
α_1
α_2
α_3

Pseudo Normal Differential Methods

Noise ratio

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_\theta(x_t,t) + \sigma_t\epsilon_t.$$

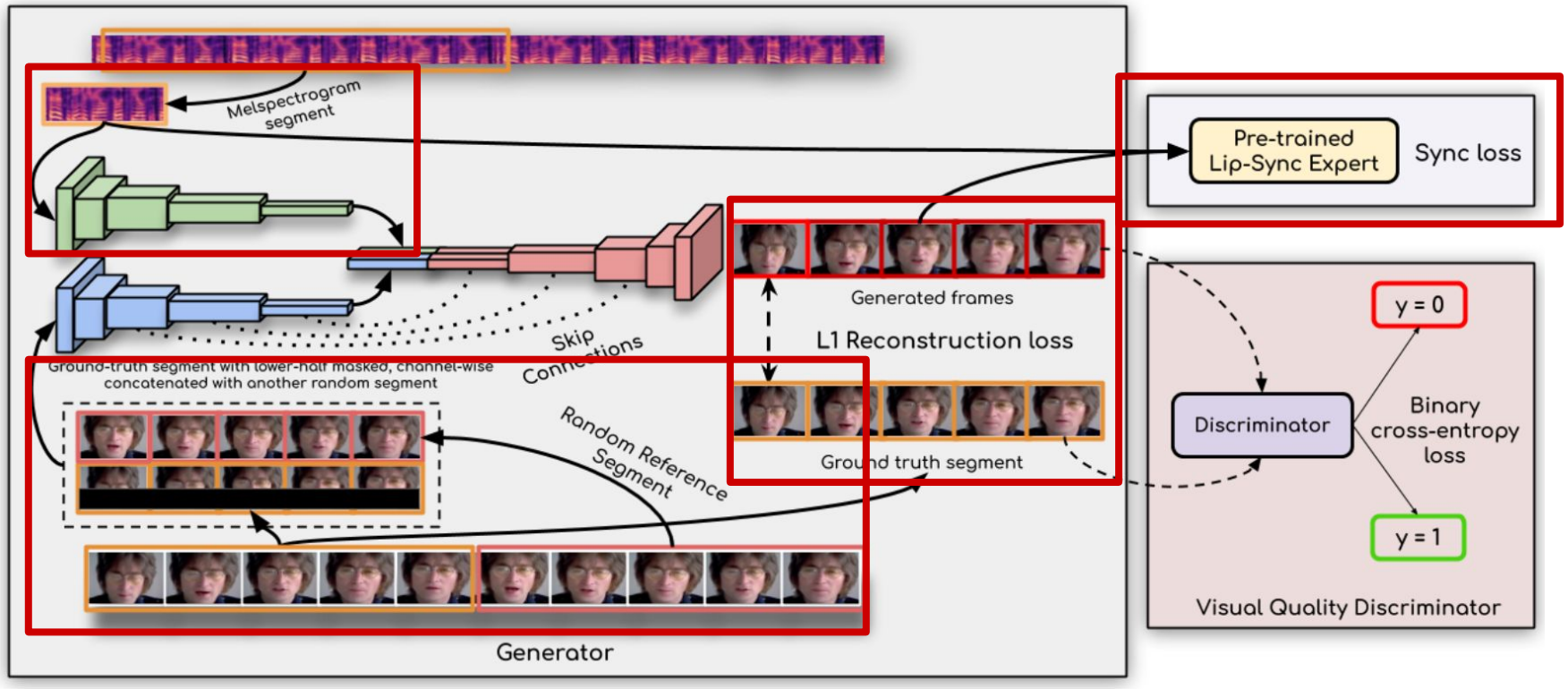
Input

$$\frac{dx}{dt} = -\bar{\alpha}'(t)\left(\frac{x(t)}{2\bar{\alpha}(t)} - \frac{\epsilon_\theta(x(t),t)}{2\bar{\alpha}(t)\sqrt{1-\bar{\alpha}(t)}}\right).$$

Differential Equation

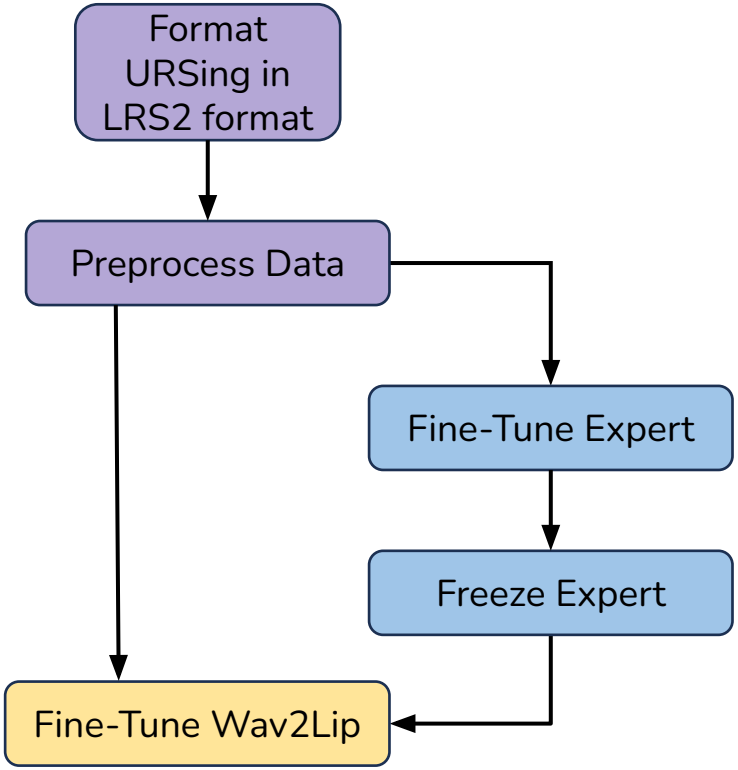
Face Synthesis

Wav2Lip


Melspectrogram segment
Pre-trained Lip-Sync Expert
Sync loss
Generated frames
L1 Reconstruction loss
Skip Connections
Ground-truth segment with lower-half masked, channel-wise concatenated with another random segment
Random Reference Segment
Ground truth segment
Generator
y = 0
Discriminator
Binary cross-entropy loss
y = 1
Visual Quality Discriminator

Fine-Tuned Wav2Lip

- URSing (Audio-Visual Solo Singing)
  - 10 audio-video songs (total length ~ 3mins each)
  - ⇒ 60 clips (30s each)
- Freeze all but last layer of Expert Discriminator.
- Fine-tune Wav2Lip
  - Freeze Expert Discriminator
  - Freeze all but last layer of Wav2Lip
    - Encoders + Visual Quality Discriminator



DeepFace

Goals:

- Improve Wav2Lip by introducing demographic information to the pipeline such as age, ethnicity and gender

Model:

- Nine layer neural network to achieve near human levels of accuracy
- Capable of Face Detection, Analysis and Verification

DeepFace Examples

{'age': 24,
'dominant_gender': 'Man',
'dominant_race': 'asian'}

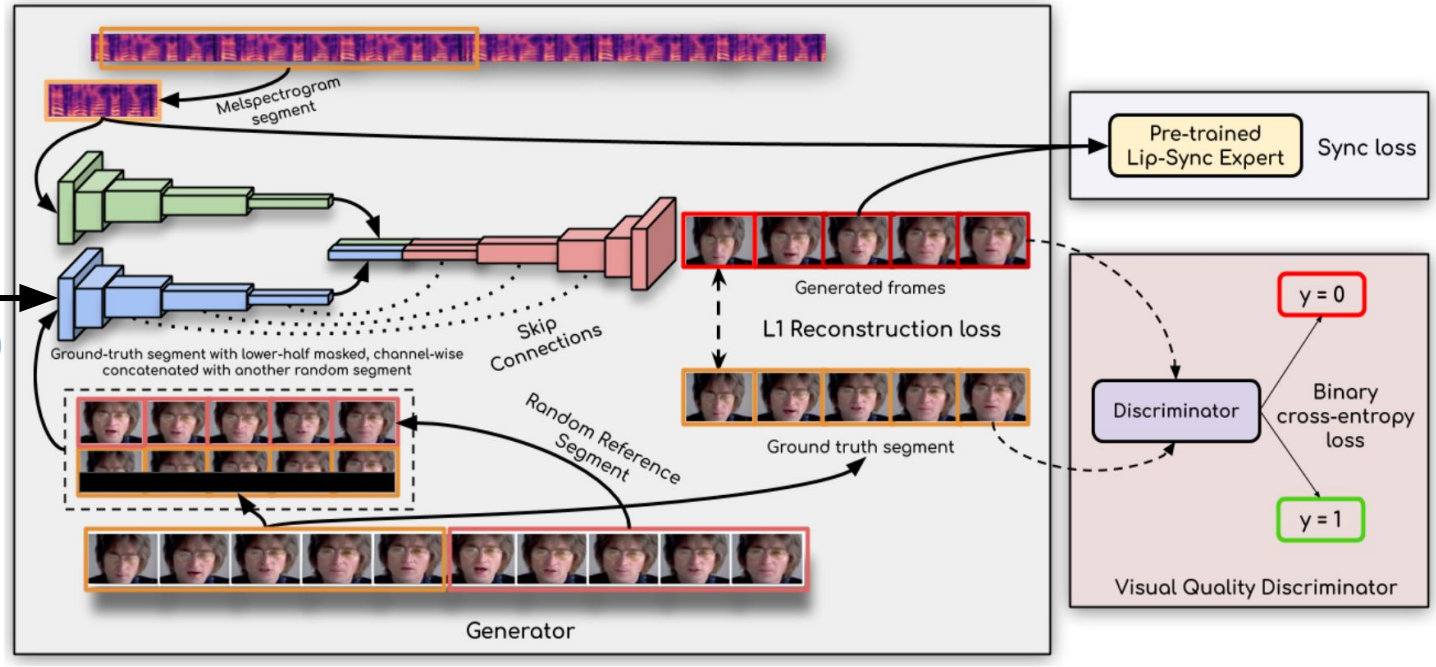
{'age': 35,
'dominant_gender': ‘Woman’,
'dominant_race': 'white'}

{'age': 28,
'dominant_gender': 'Woman',
'dominant_race': 'black'}

DeepFace Augmentation

DeepFace Information (Age, Ethnicity, Gender)

+



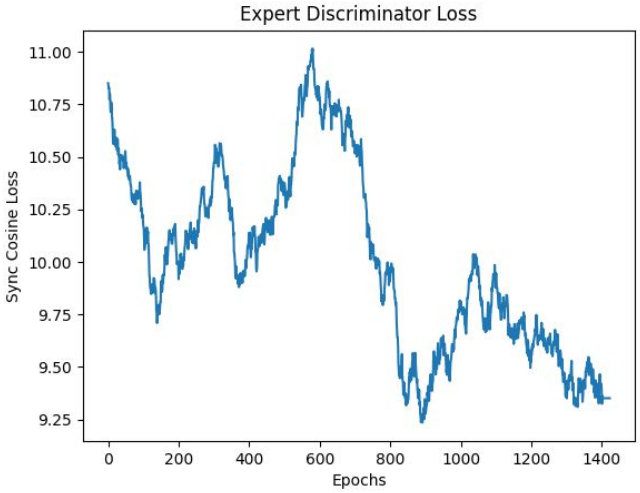
Results

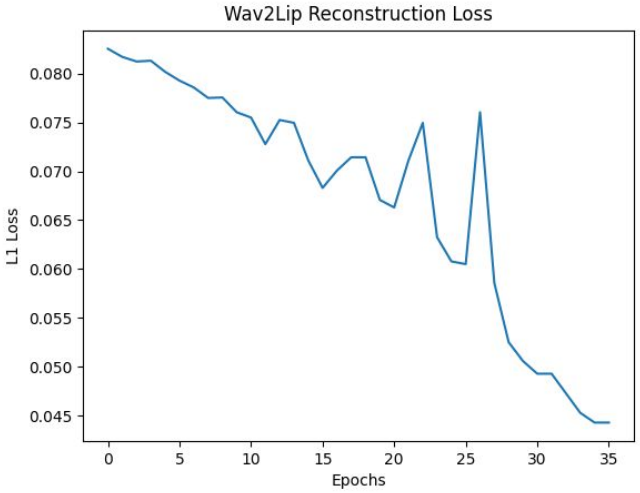
Results - DiffSinger

- Signal to Noise Ratio for Vocoder
  - HifiGAN: `-0.00022`
  - BigVGAN: `-0.00044`

- PNDM inference speedup
  - `40`x in `25` steps vs `1000` steps

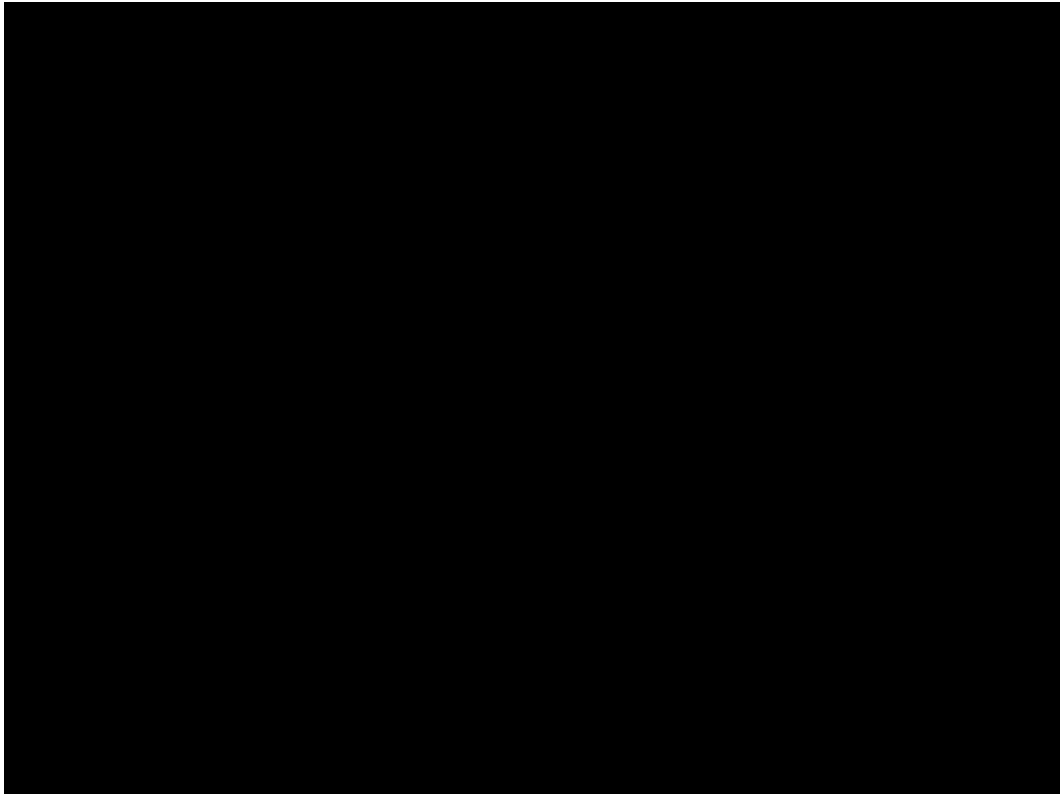
Results - Wav2Lip Fine-tuning





Trained on **three RTX3090**

Results - Wav2Lip DeepFace Augmentation



Demo

References

[1] K R Prajwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, C V Jawahar. (2020). **A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild.**

[2] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, Fei Wang. (2022). **SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation**

[3] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, & Zhou Zhao. (2022). **DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism.**

[4] Luping Liu, Yi Ren, Zhijie Lin, & Zhou Zhao. (2022). **Pseudo Numerical Methods for Diffusion Models on Manifolds.**

[5] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, & Sungroh Yoon. (2023). **BigVGAN: A Universal Neural Vocoder with Large-Scale Training.**

Thank You!

Future Work

Possible Improvements

- Improving the DeepFace augmentation to seamlessly integrate the information into the input face data, allowing us to further improve the system and remove the artifacts.