ICASSP 2023

Free-View Expressive Talking Head Video Editing

Yuantian Huang, Satoshi Iizuka, Kazuhiro Fukui

University of Tsukuba

1 Introduction

➢Our Goal

- The goal of our research is to **edit** the talking head in a video with **multiple attributes in full frames**, including head pose, facial emotion, and eye blink.

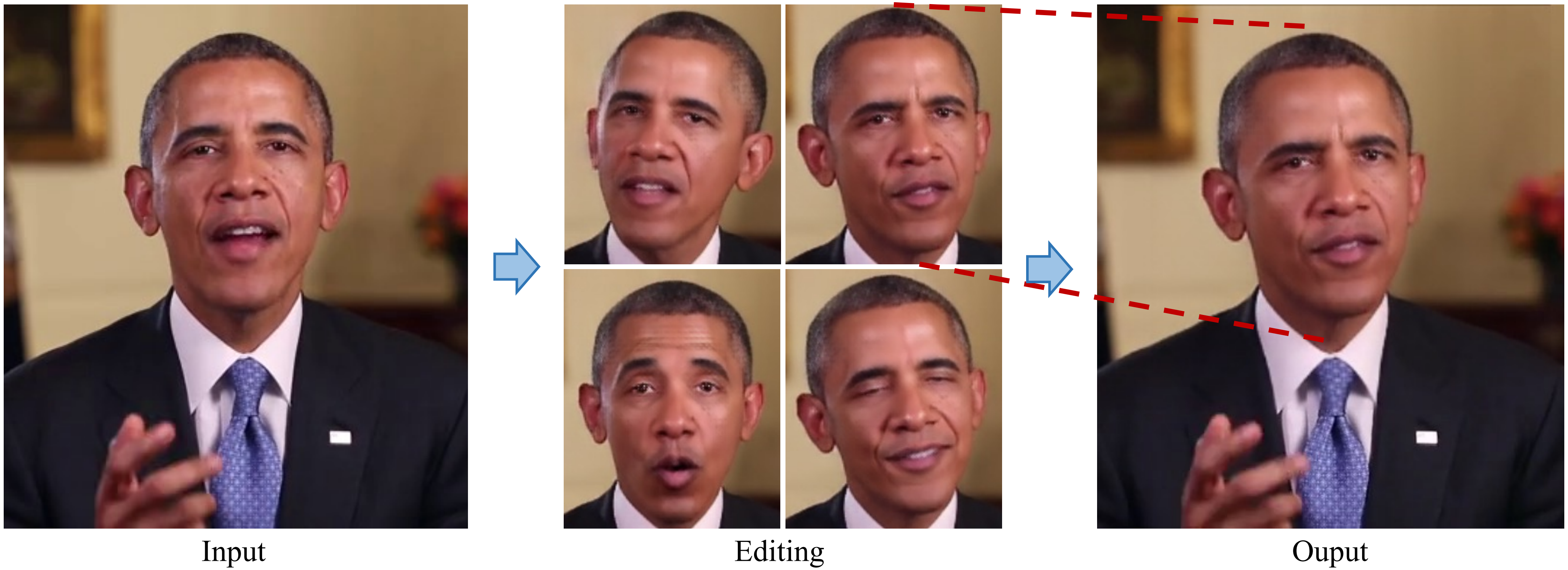


Figure 1: Teaser.

➢Previous Research

- **Previous Research** → Audio-driven Talking Head Generation
- **Proposed Method** → **Editing instead of Generation**

Unlike previous approaches that mainly focus on generating talking head videos, our proposed method is able to edit the talking head and restore it back to the full frames, which supports a broader range of applications.

Methods	Lip Sync	Full-Frame Editing	Head Pose	Emotion	Eye Blink
Wav2Lip [1]	✓	✓	Copy source		
EAMM [2]	✓		Copy source	✓	
PC-AVS [3]	✓		Reference required		
GC-AVT [4]	✓		Reference required	✓	
Ours	✓	✓	Free-view	✓	✓

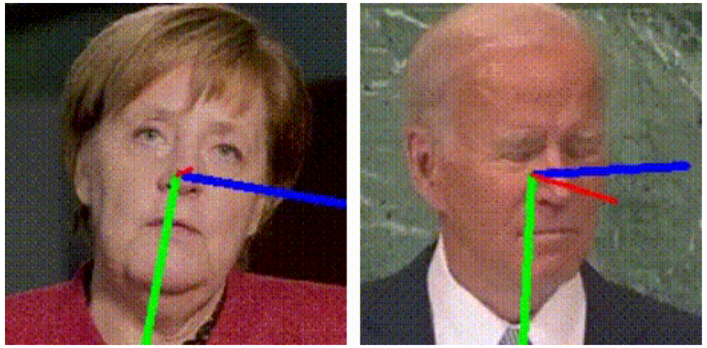
Table 1: Comparison with Previous Work.

2 Proposed Framework

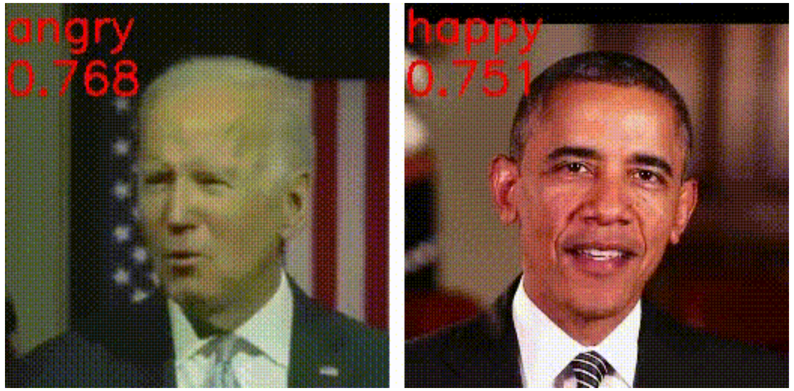
➢Model Input

- **Typical Input** for Audio-driven Talking Head Generation
  - **Input frames and target frames** with black-masked faces.
  - **Audio Source** that are corresponding to target frames.

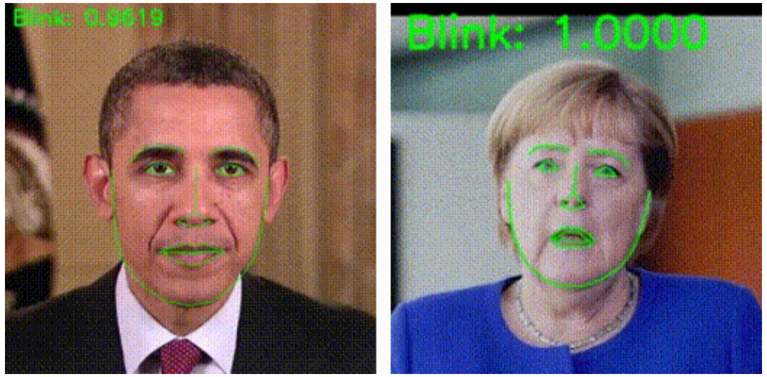
- **Attribute Input** of our proposed method
  - **Head pose** code that represents angles of the talking head (yaw, pitch,
  - **Emotion** code that indicates seven emotion categories.
  - **Eye blink** code that is related to eye openness.



Head Pose



Emotion



Eye Blink

➢Model Architecture

1. A set of encoders that embed inputs together and feed them into the generator.
All embed inputs are resampled to be the same length as the audio source (Mel Spectrogram).

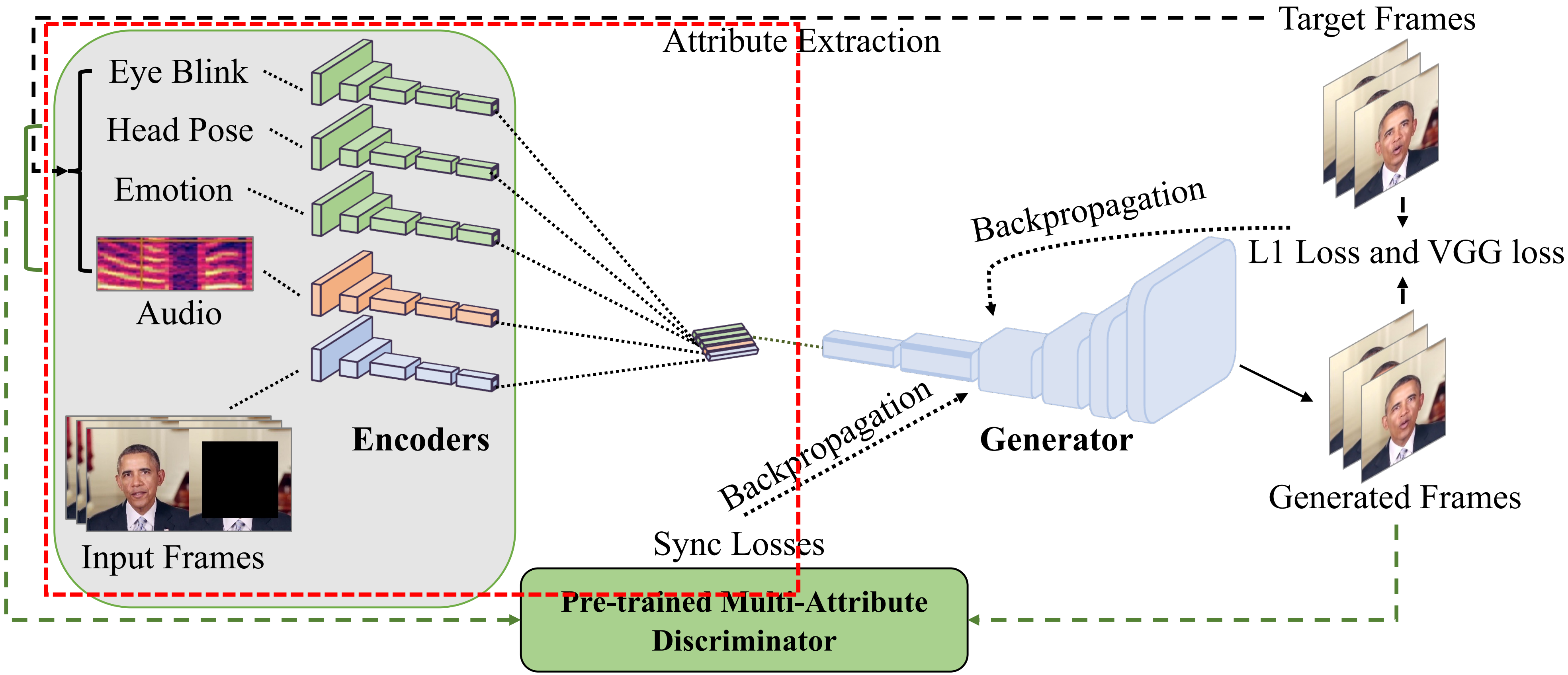


Figure 2: Structure of Proposed Framework.

➢Model Architecture

2. A generator that generates frames corresponding to the input.
L1 loss and VGG perceptual loss are used to improve the overall generation quality.

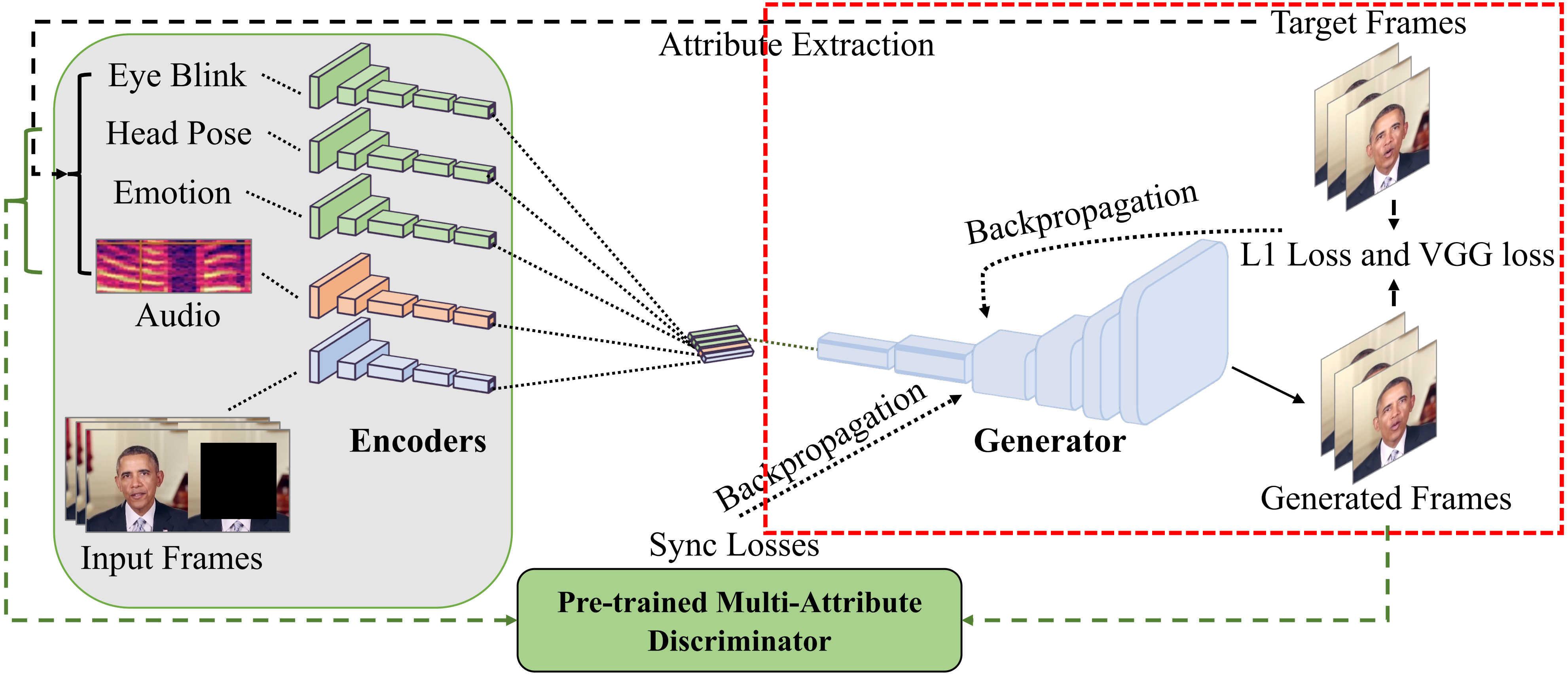


Figure 2: Structure of Proposed Framework.

➢Model Architecture

3. A pretrained multi-attribute discriminator that calculates a set of synchronization losses between generated frames and input attributes to enforce attribute-visual synchronization.

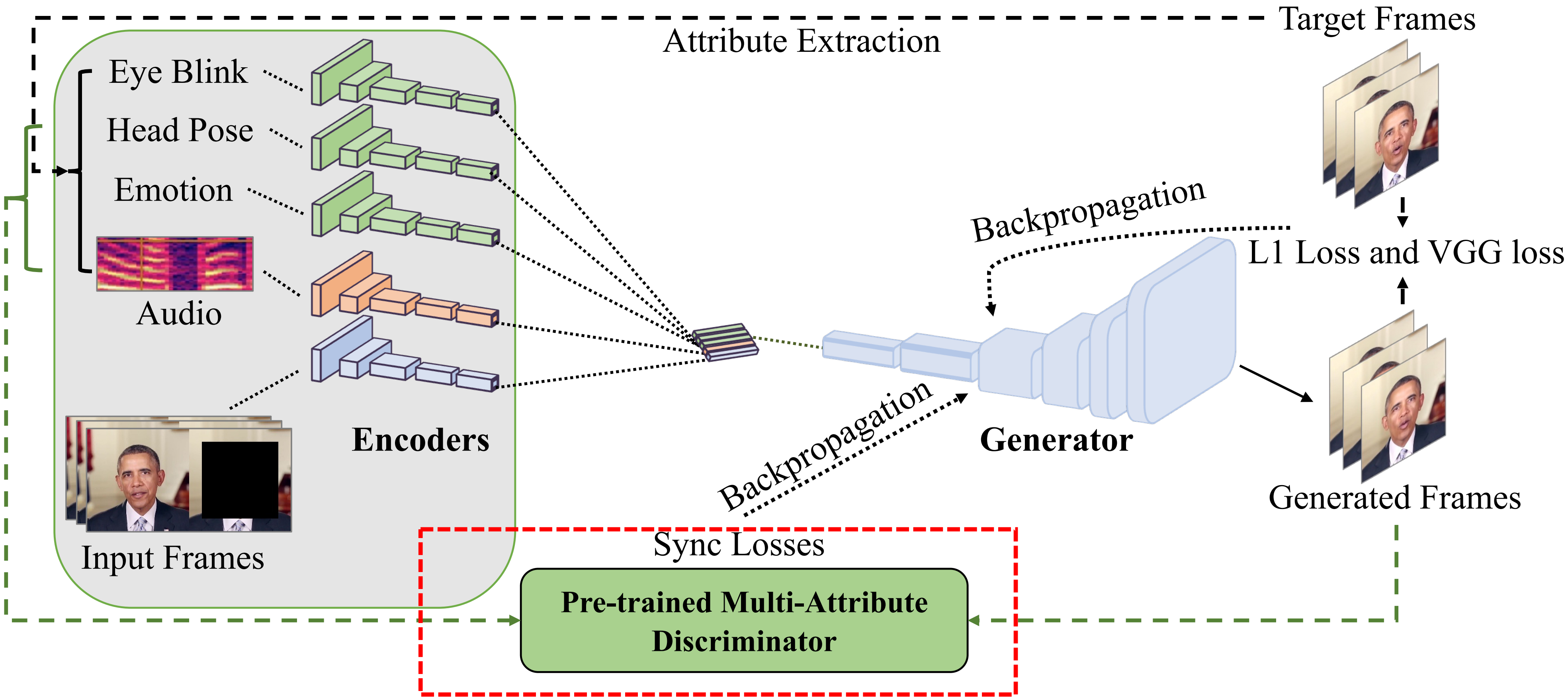


Figure 2: Structure of Proposed Framework.

➢Loss Function

- **Synchronization loss:**

We minimize the *L2* distance between embedding features that are extracted from the target frames and those from other inputs. So that we are able to enforce a synchronization between generated frames and all the input attributes simultaneously.

$$L_{sync} = L_{audio} + L_{pose} + L_{emotion} \qquad (1)$$

- **Full objective:**

$$L_{total} = L_{l1} + \lambda_v L_{perceptual} + \lambda_s L_{sync} + \lambda_b L_{blink} \qquad (2)$$

3 Results

➢ Attributes Editing (Without Reference Input)

- Our proposed framework is able to fully control three different types of facial attributes, including head pose, facial emotion, and eye blink.

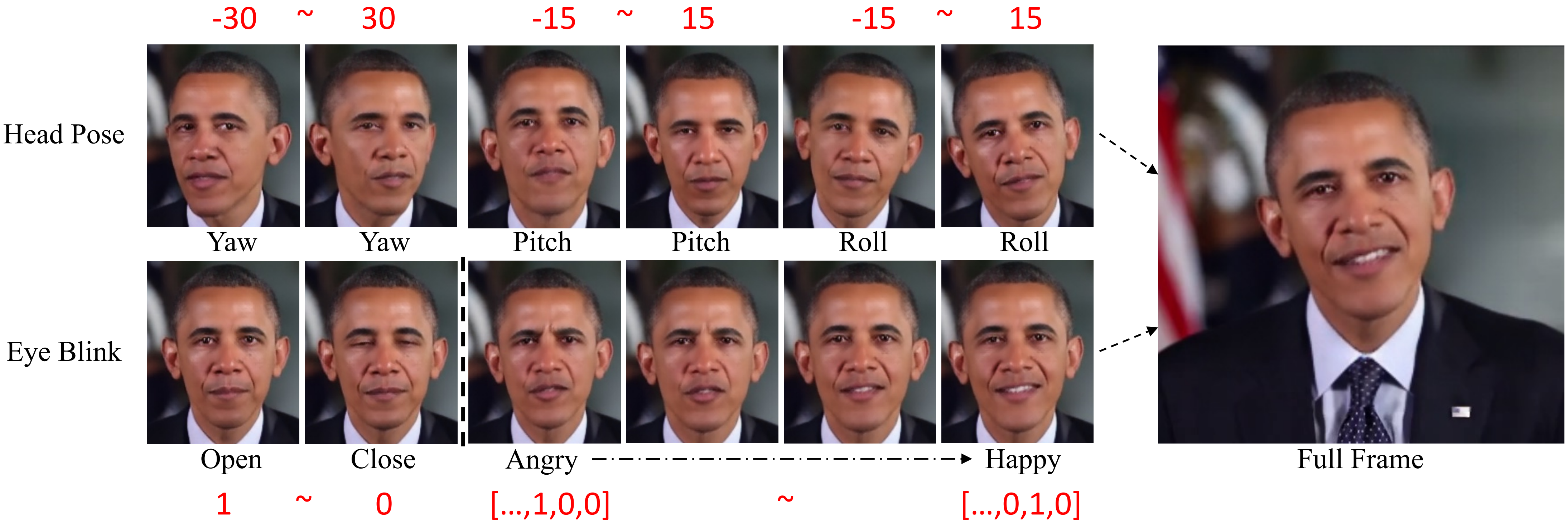


Figure 3: Examples of Attributes Editing.

➤Visual Quality Comparison (With Reference Input)

- **Wav2Lip** fails to change any attributes since it only accepts audio input without any additional attributes.
- **PC-AVS** can change head poses properly. However, it can not give these changes back to the original frames.
- Baseline models also fail to change emotion and eye blink corresponding to the reference video.

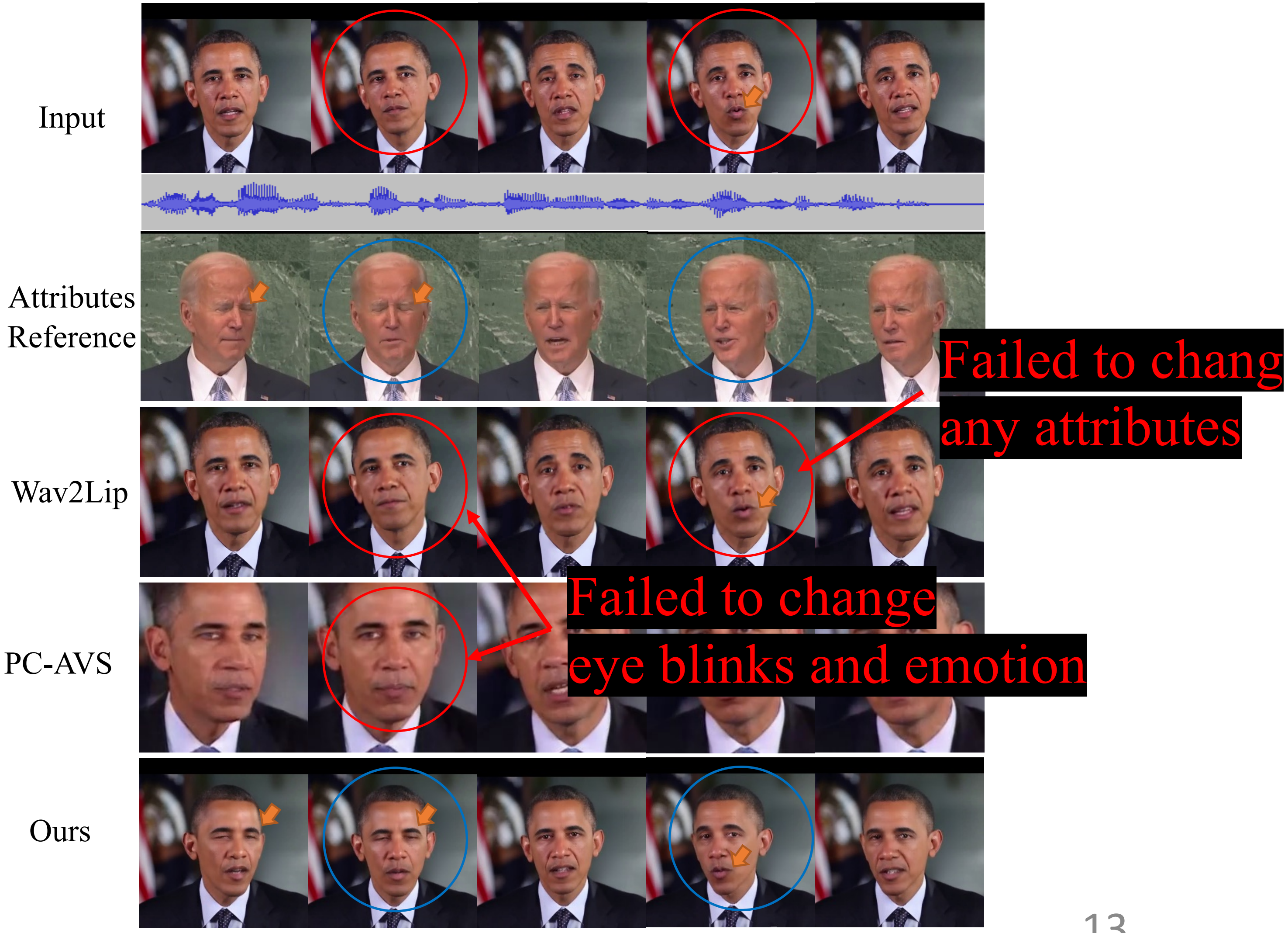
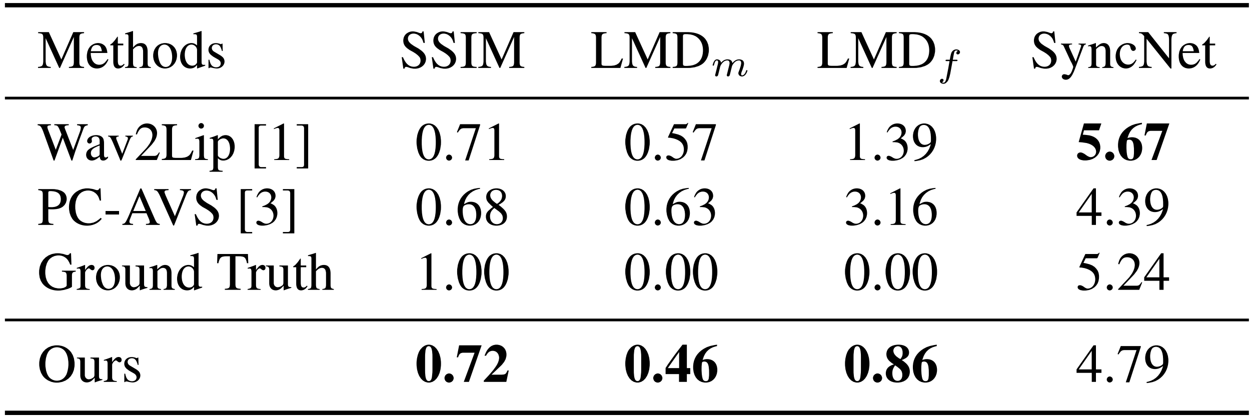

Input
Attributes Reference
Wav2Lip
PC-AVS
Ours
Failed to chang any attributes
Failed to change eye blinks and emotion

Figure 4: Visual Quality Comparisons .

➢Quantitative Comparison

- Metrics:
  - **SSIM** (structural similarity index measure) to evaluate the **video generation quality**.
  - **SyncNet scores** and **LMD_m** (landmark distance on the mouth) to evaluate **mouth movements.**
  - **LMD_f** (landmark distance on the face) to evaluate how well the facial expression and eye blink fit the desired results.



Methods	SSIM	LMD$_m$	LMD$_f$	SyncNet
Wav2Lip [1]	0.71	0.57	1.39	**5.67**
PC-AVS [3]	0.68	0.63	3.16	4.39
Ground Truth	1.00	0.00	0.00	5.24
Ours	**0.72**	**0.46**	**0.86**	4.79

Figure 5: Quantitative Comparison.

➢Video Example

➢Conclusion

- **Free-View Expressive Talking Head Video Editing**

1. A reconstruction-based generator that can generate talking heads fitting to the original frame while corresponding to **freely controllable attributes**, including head pose, facial emotion, and eye blink.
2. A multi-attribute discriminator that enforces attribute-visual synchronization.
3. By combining these two modules, our proposed model **can edit talking head videos in full frames** on multiple attributes with or without references, which is one of the earliest attempts at this task.

o **The main limitation** of our proposed model is that it may not be able to generate satisfactory results when training with multiple identities. We expect a general model that includes identity learning in our future research.

Thank You For Listening!