Talking Face Generation via Learning Semantic and Temporal Synchronous Landmarks

Aihua Zheng[1], Feixia Zhu[1], Hao Zhu[1], Mandi Luo[2,3] and Ran He[2,3,4]

[1]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation,

School of Computer Science and Technology, Anhui University, Heifei, China

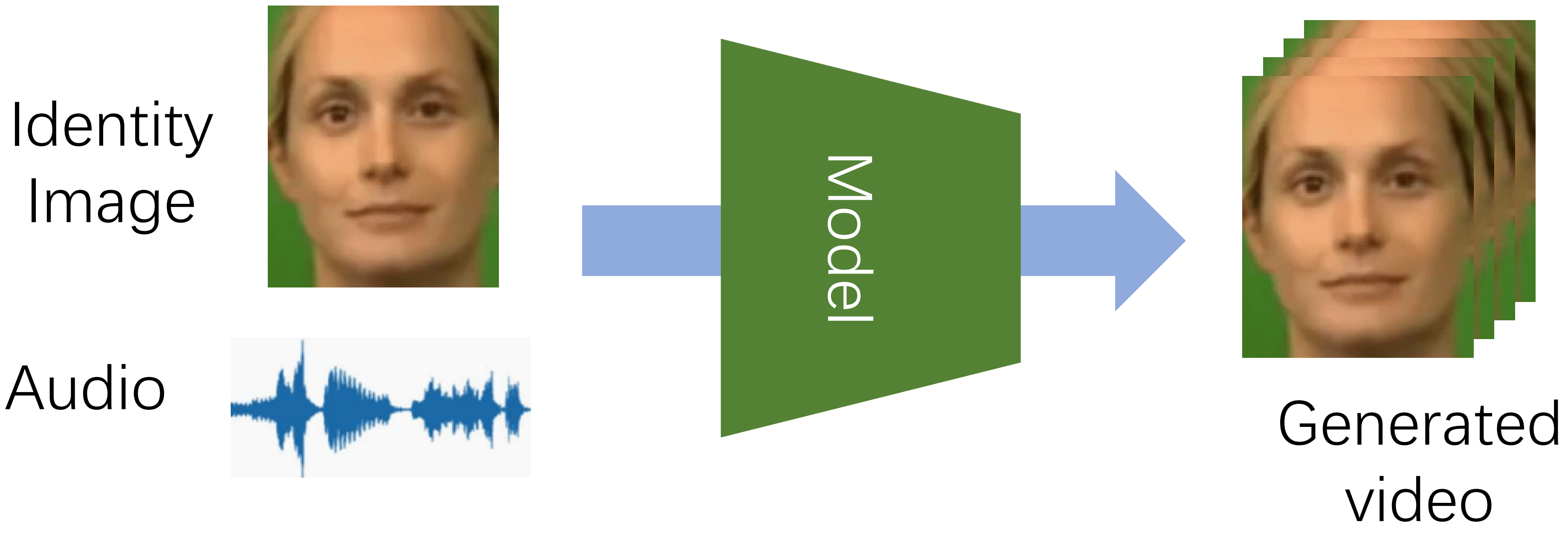
[2]University of Chinese Academy of Sciences, Beijing, China

[3]Center for Research on Intelligent Perception and Computing (CRIPAC)

[4]National Laboratory of Pattern Recognition (NLPR), CASIA, Beijing, China

ICPR 2020

Task Definition



Identity Image

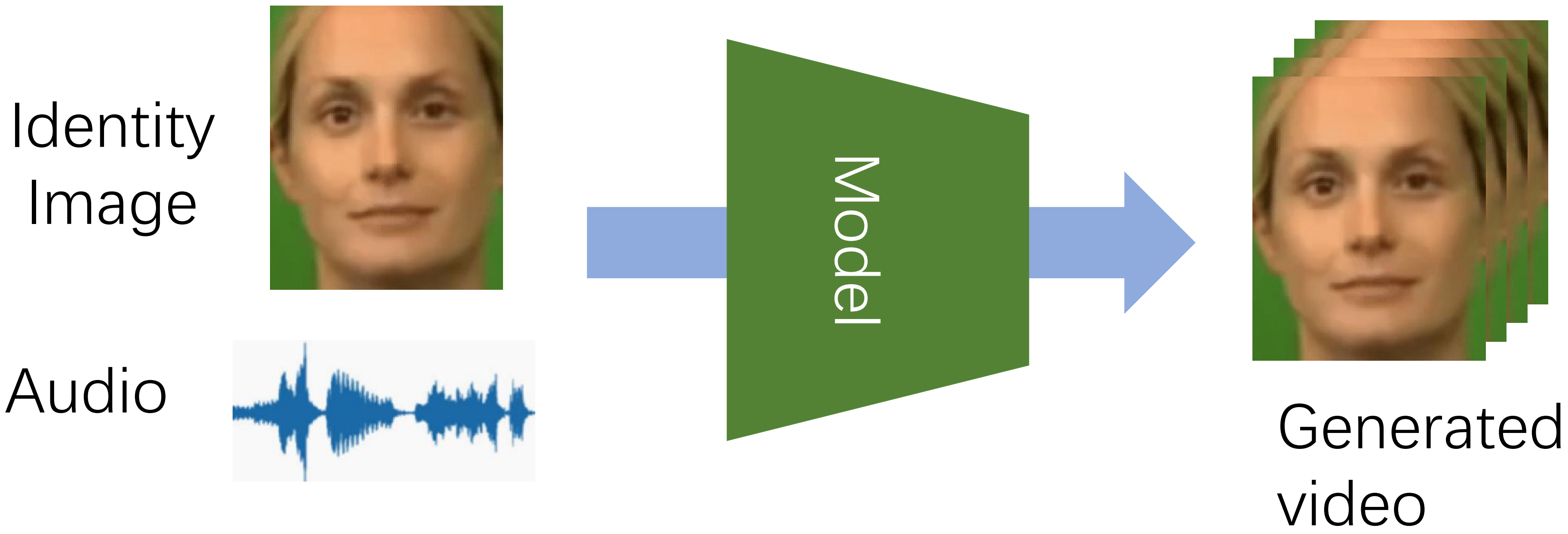
Audio

Model

Generated video

- Aims:
  - Realistic talking face video with lip synchronization
  - Smooth facial motion

Task Definition



Identity Image

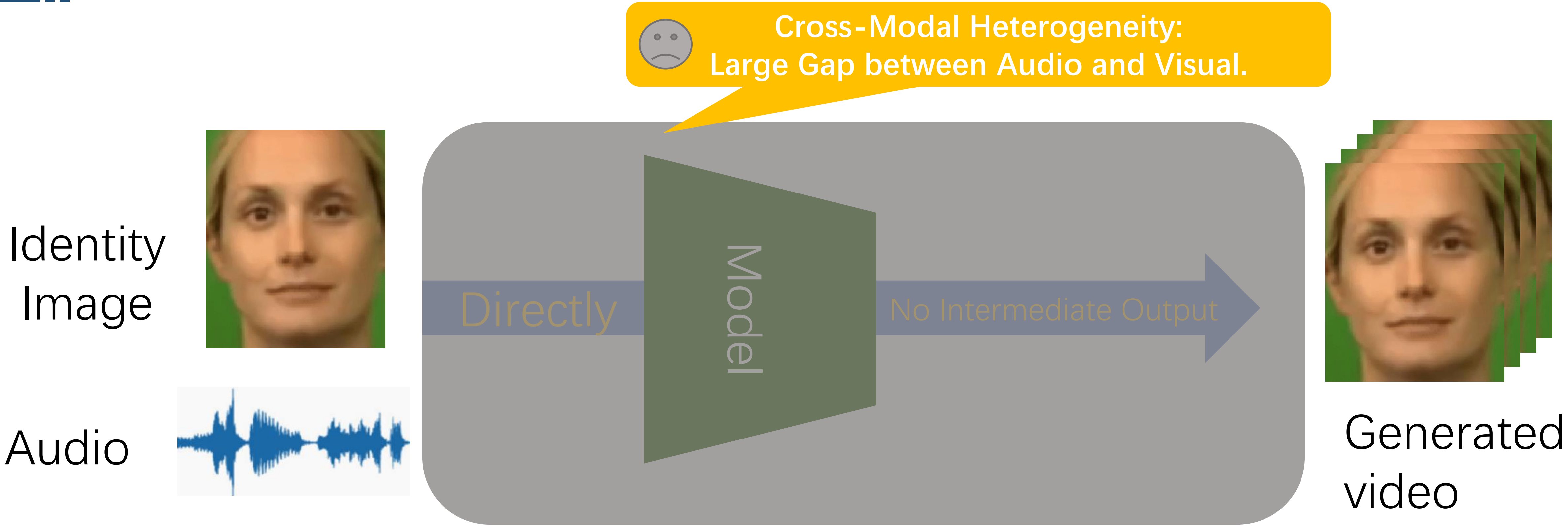
Audio

Model

Generated video

- Applications: virtual computer games, speech comprehension, and teleconferencing and so on.

Task Definition



Cross-Modal Heterogeneity:
Large Gap between Audio and Visual.

Identity Image

Audio

Directly

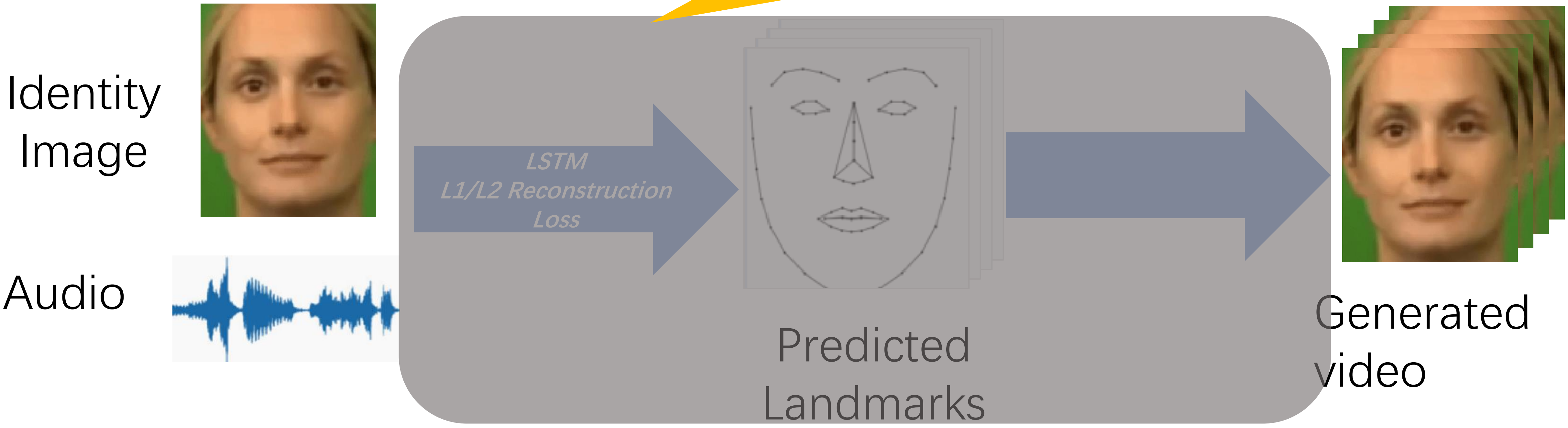
Model

No Intermediate Output

Generated video

- Previous works: generate talking face video directly.

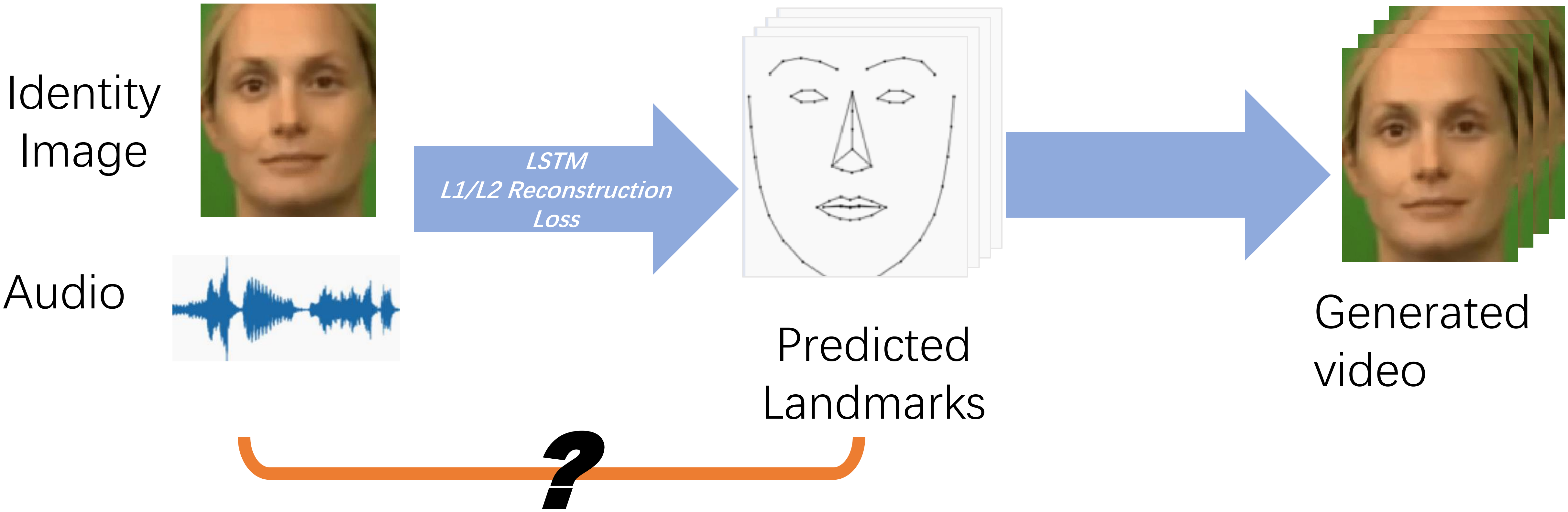
Task Definition



There is still challenge to predict landmarks from audio.

- Two-stage Strategy: gradually close the gap by predicted landmarks.

Motivation



Identity Image

Audio

LSTM
L1/L2 Reconstruction Loss

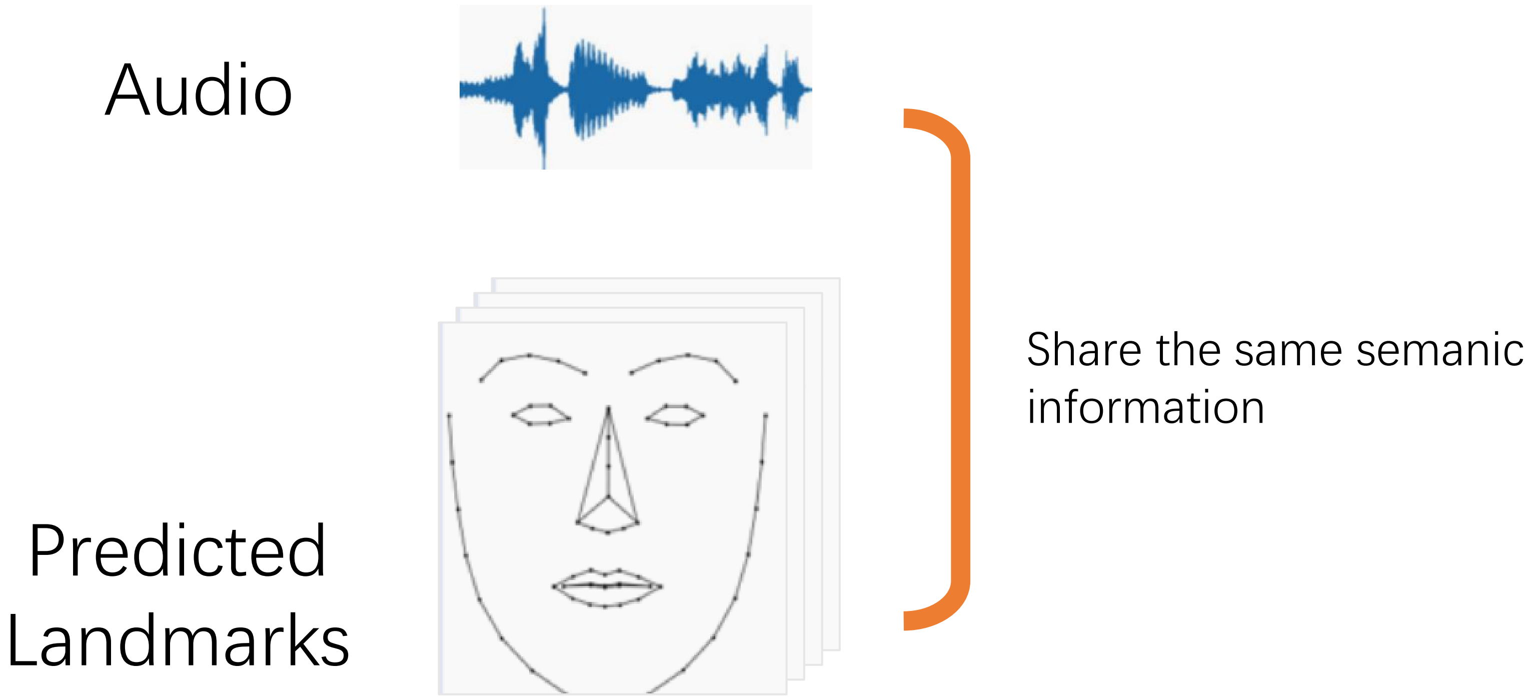
Predicted Landmarks

Generated video

?

- How to learn more reasonable and synchronous landmarks from audio to guide talking face generation?

Motivation



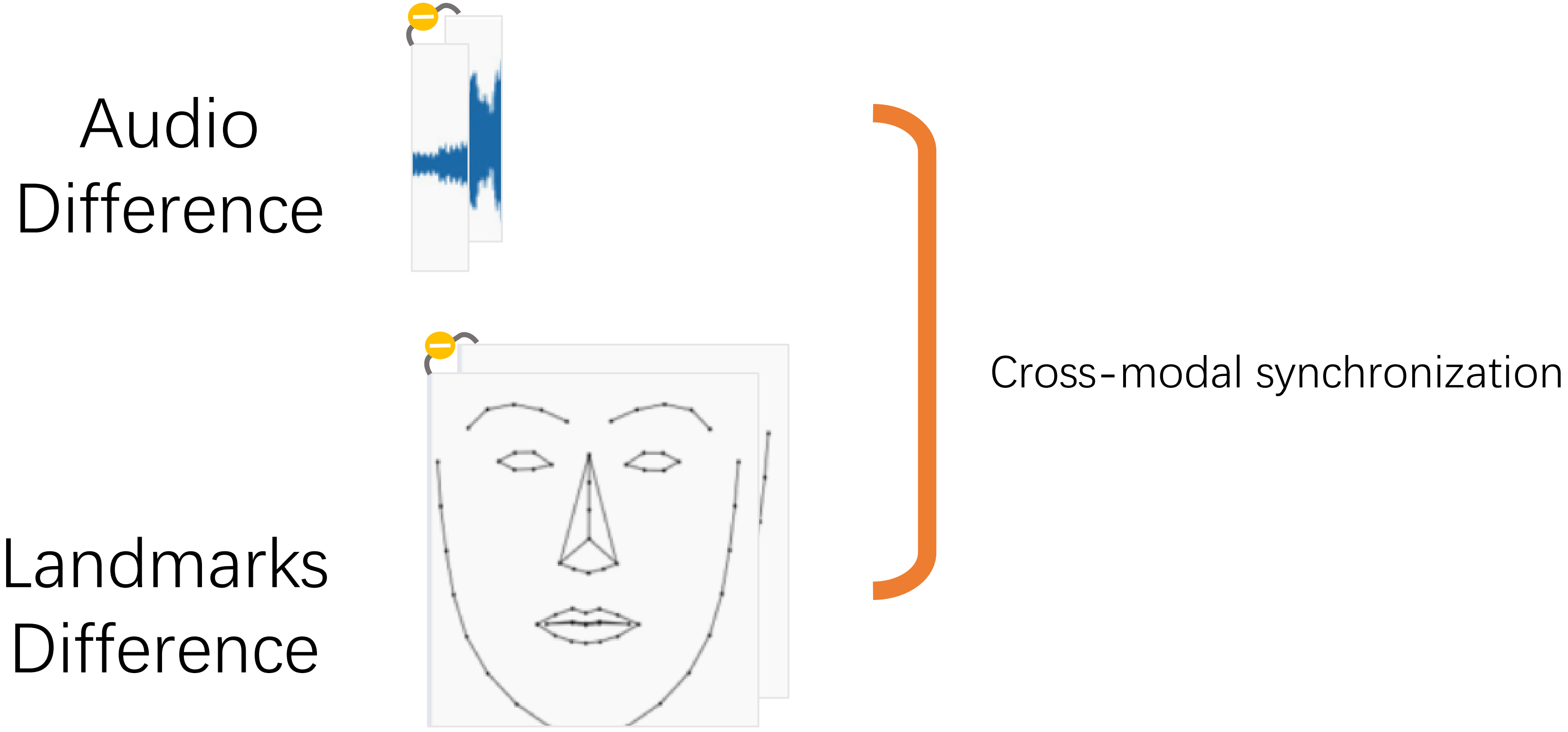
Audio

Predicted Landmarks

Share the same semanic information

- The words play a important role to bridge the audio and landmarks modalities.

Motivation



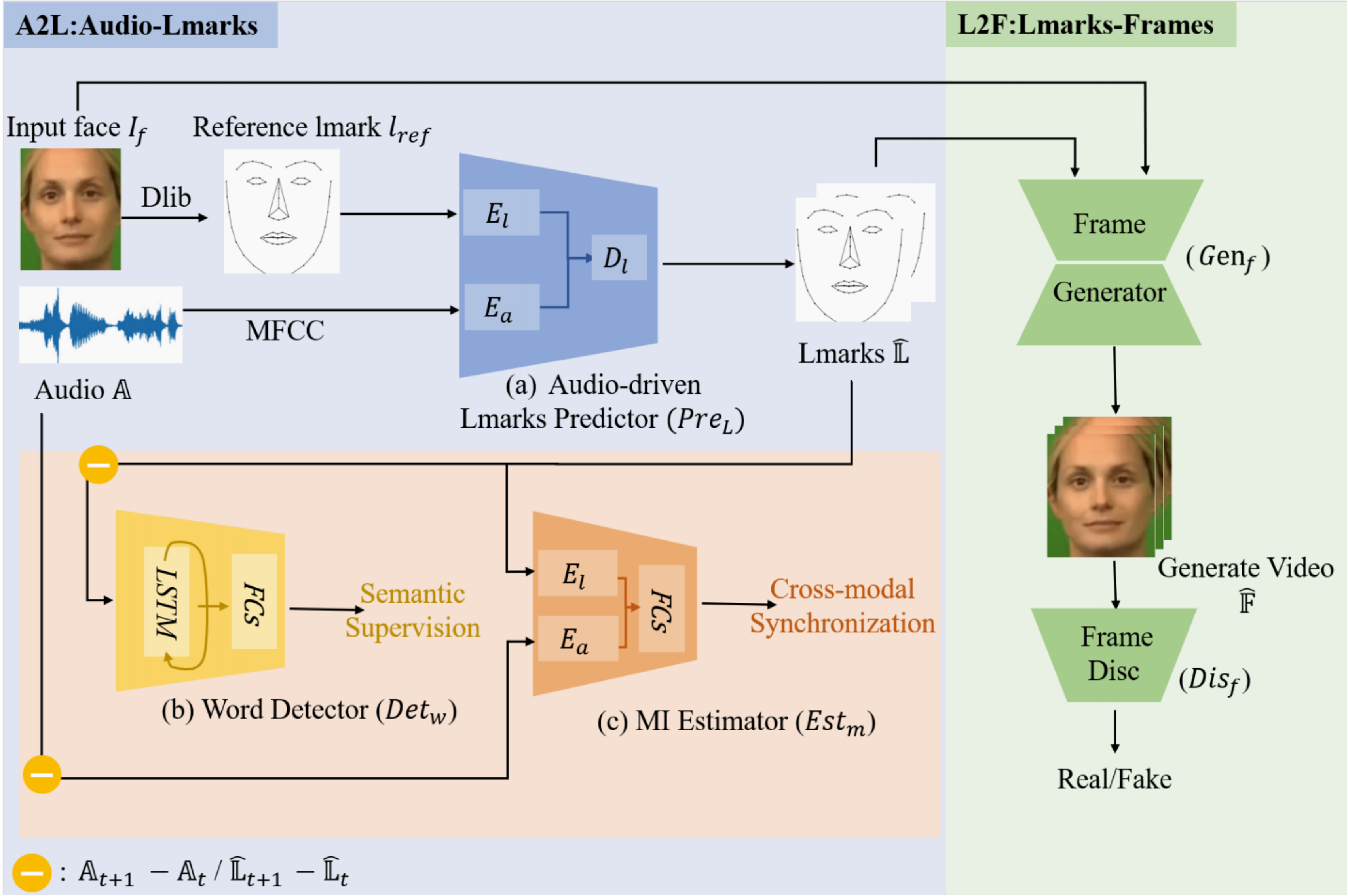
Audio Difference

Landmarks Difference

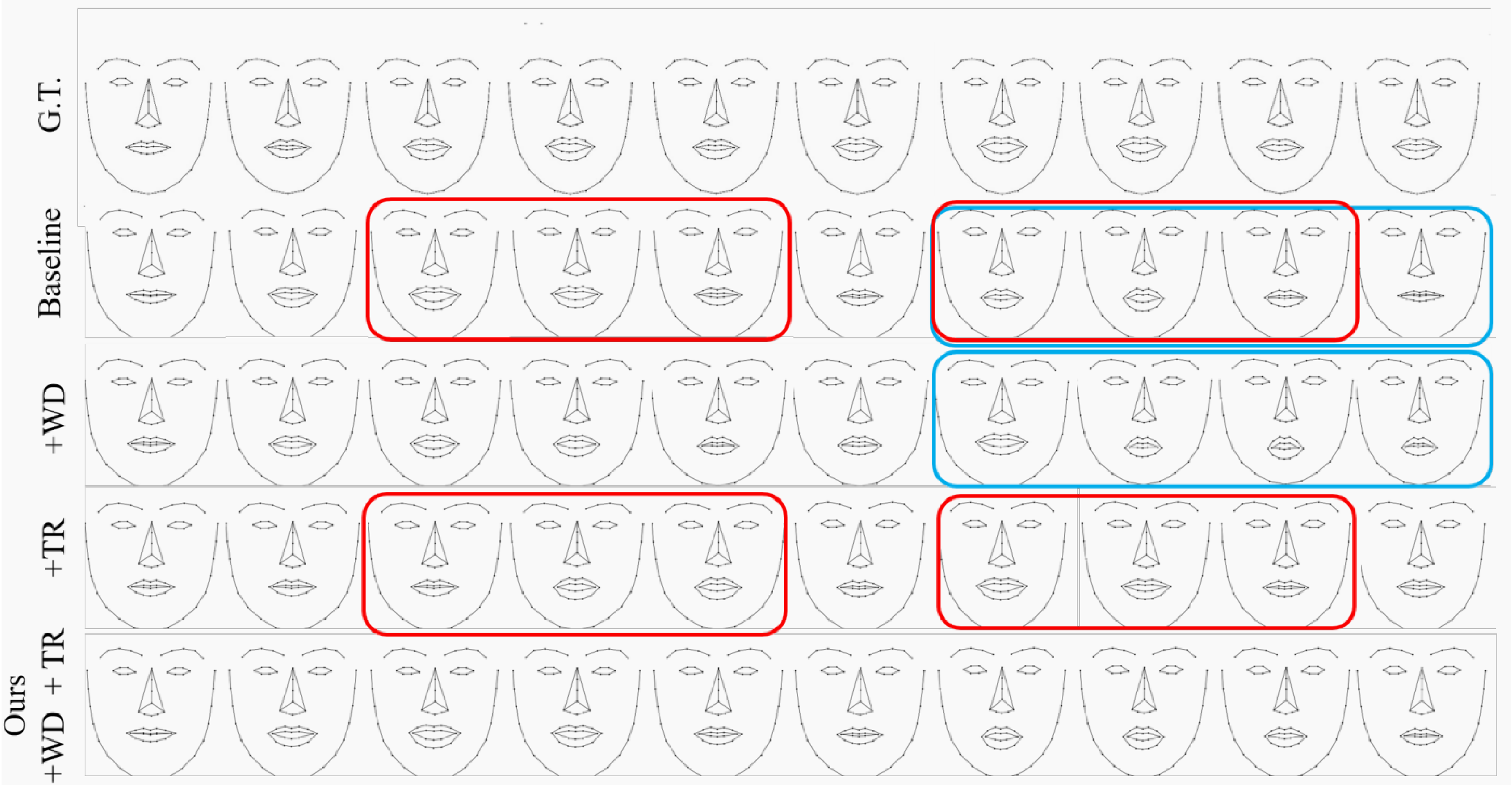
Cross-modal synchronization

- Temporal consistency is important for the authenticity of the smooth transition between frames in sequence.

Approach


A2L:Audio-Lmarks
L2F:Lmarks-Frames
Input face I_f
Reference lmark l_{ref}
Dlib
E_l
D_l
E_a
MFCC
Audio $\mathbb{A}$
Lmarks $\hat{\mathbb{L}}$
(a) Audio-driven Lmarks Predictor (Pre_L)
Frame Generator
(Gen_f)
Generate Video $\hat{\mathbb{F}}$
Frame Disc
(Dis_f)
Real/Fake
LSTM
FCs
Semantic Supervision
(b) Word Detector (Det_w)
E_l
E_a
FCs
Cross-modal Synchronization
(c) MI Estimator (Est_m)
: $\mathbb{A}_{t+1} - \mathbb{A}_t$ / $\hat{\mathbb{L}}_{t+1} - \hat{\mathbb{L}}_t$

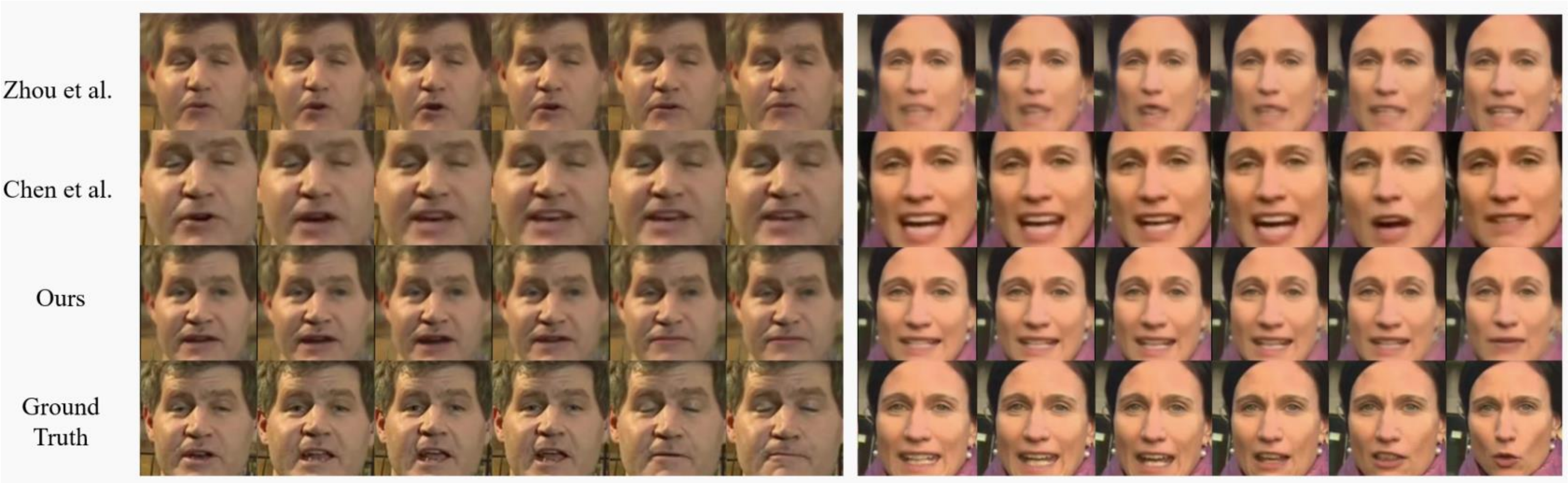
Experiments



(a) The landmarks of ‘BETTER’

- Blue boxes: the difference in word semantic pronunciation.
- Red boxes: the difference between landmarks in temporal consistency.

Experiments



- The examples of generated talking faces from test set of LRW dataset.

Thanks for watching