SpeechForensics: Audio-Visual Speech Representation Learning for Face Forgery Detection

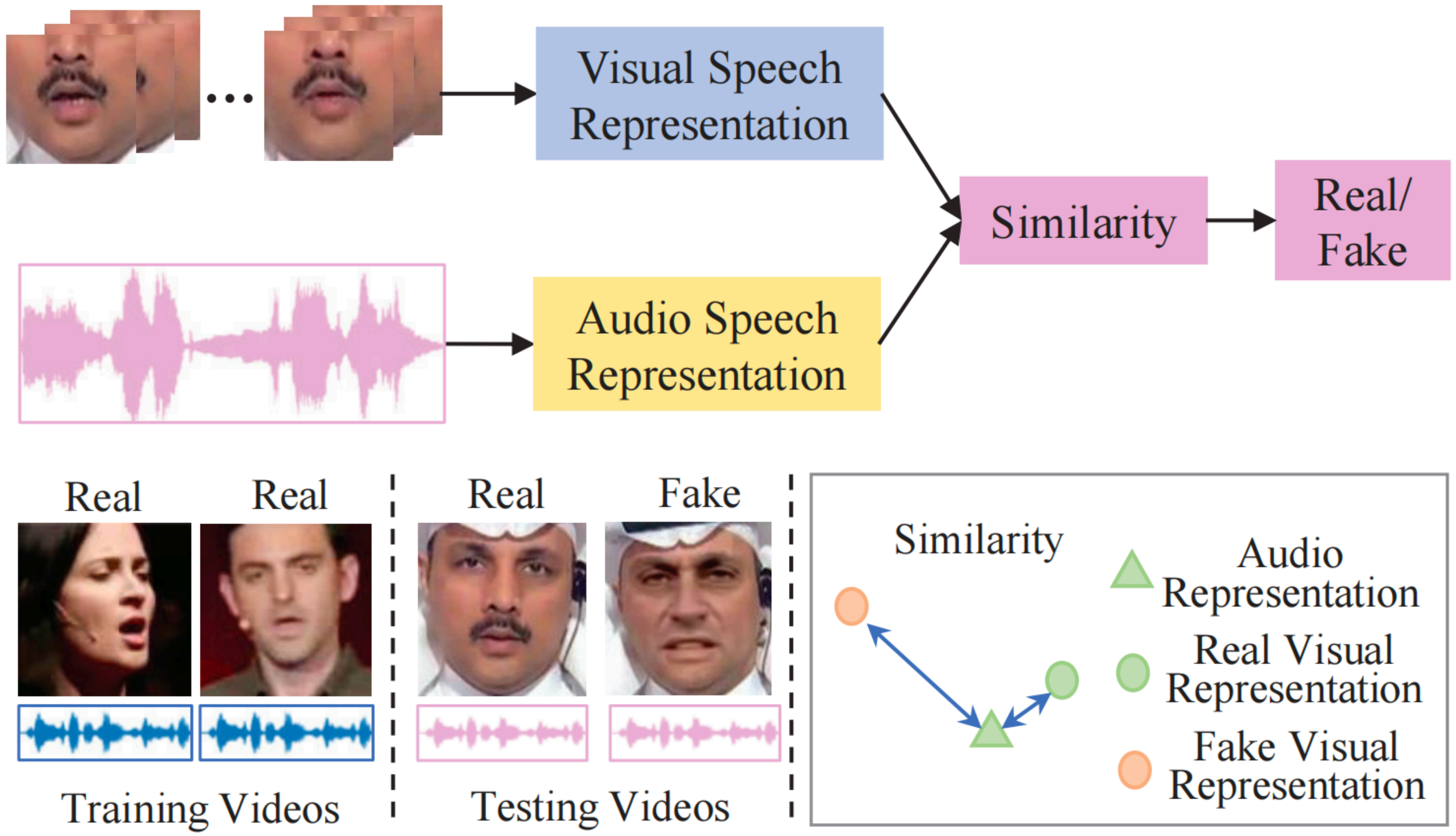
Yachao Liang, Min Yu, Gang Li, Jianguo Jiang, Boquan Li
Feng Yu, Ning Zhang, Xiang Meng, Weiqing Huang

中國科學院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

1. Existing facial forgeries are typically performed at the frame level, which can lead to **temporal inconsistencies**.
2. The movement of the lips is strongly correlated with the spoken words of the person, while the audio signal in the video can accurately capture them.
3. In real videos, the speech contents conveyed by the mouth movements should be consistent with the counterparts extracted from audio signals.



Method

Training

Inference

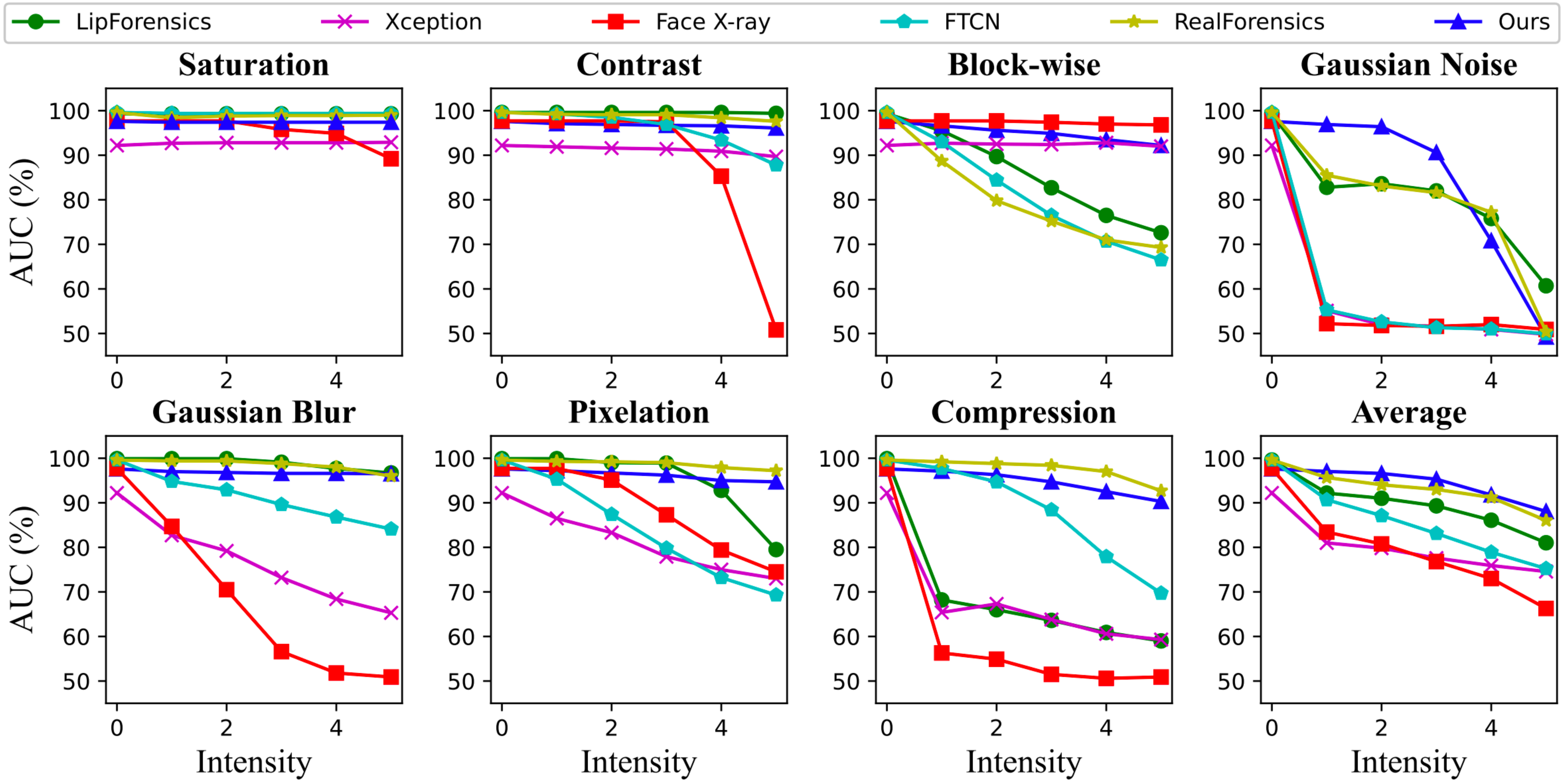
Generalization

Thanks to the self-supervised learning manner, our method shows promising generalization, especially under the cross-dataset setting.

Table 2: **Cross-dataset generalization.** Video-level AUC (%) on FakeAVCeleb and KoDF. We report the results of every categories of FakeAVCeleb, and the overall performance on it is reported in **Overall**. The average performance over two datasets is reported in **Avg**.

	Method	FakeAVCeleb						KoDF	**Avg**
		FS	FSGAN	WL	FS-WL	FSGAN-WL	**Overall**		
Supervised	Xception [52]	67.0	62.5	59.7	57.2	68.0	61.6	77.7	69.7
	Patch-based [12]	97.4	80.5	78.9	93.8	87.8	83.6	83.9	83.8
	Face X-ray [41]	89.9	85.4	69.5	84.4	87.6	78.4	83.0	80.7
	LipForensics [28]	89.5	96.4	85.6	87.2	95.8	89.8	59.6	74.7
	FTCN[67]	89.3	79.9	80.6	85.2	86.1	82.3	76.5	79.4
	RealForensics [27]	**98.1**	**100.**	81.0	94.7	99.2	90.2	84.3	87.3
Unsupervised	AVAD* [22]	52.8	53.9	93.9	95.8	94.3	85.0	58.0	71.5
	SpeechForensics-Local	69.3	85.4	0.10	0.08	0.08	19.0	48.3	33.7
	SpeechForensics (ours)	93.9	96.0	**100.**	**99.9**	**99.9**	**99.0**	**91.7**	**95.4**

Robustness


LipForensics
Xception
Face X-ray
FTCN
RealForensics
Ours
Saturation
Contrast
Block-wise
Gaussian Noise
Gaussian Blur
Pixelation
Compression
Average
AUC (%)
Intensity
100
90
80
70
60
50
0
2
4

Interpretability

Detection can also be conducted with differences between transcriptions of lip movements and audio signals.

But this requires speech recognition models to be able to recognize different languages!

Ablation Study

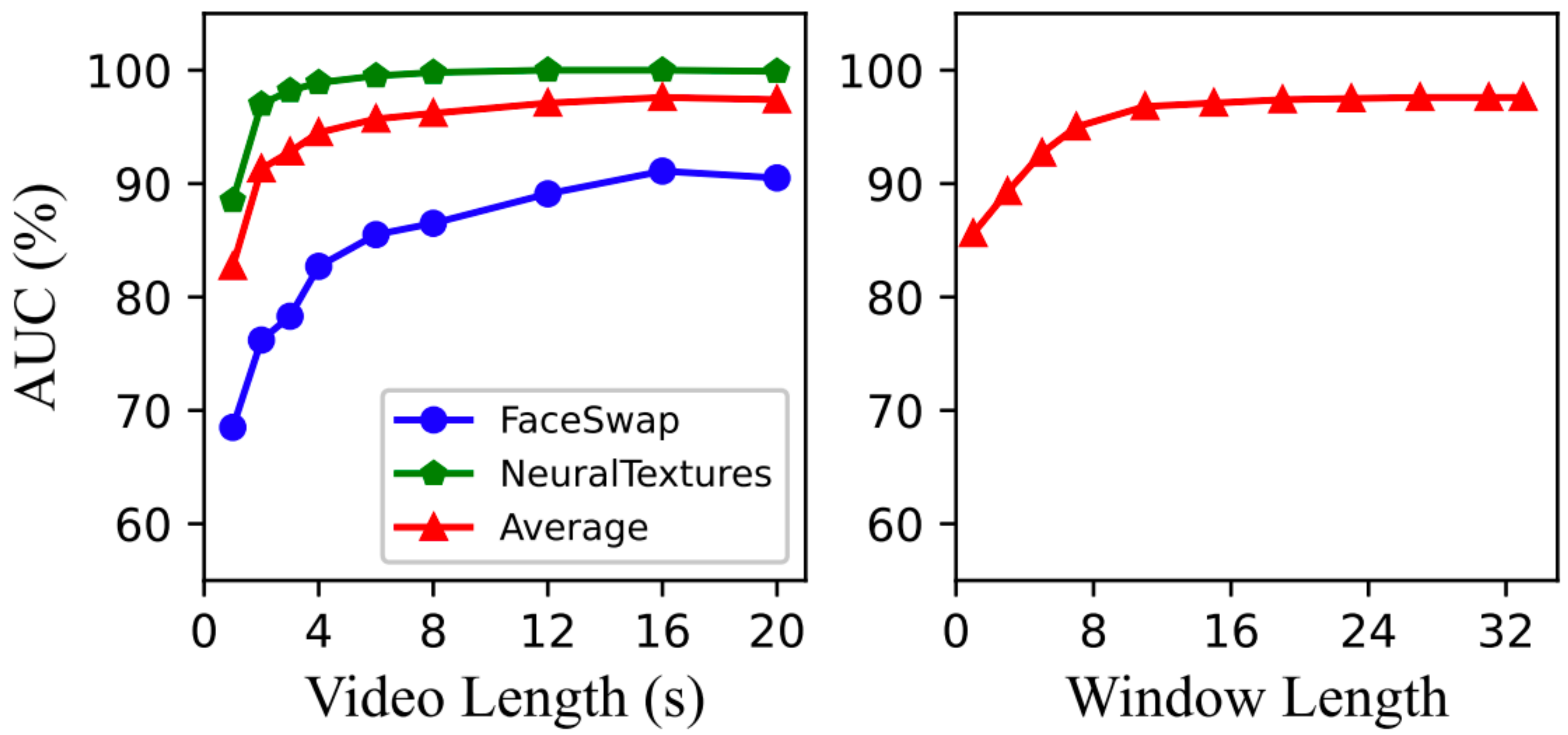


Figure 5: **Influence of video length and sliding-window length.** We evaluate the performance of our method conditioned on different input lengths and sliding-window lengths.

Table 4: **Effect of different models and time offset assumptions.** We report the performance of models with different architectures and training datasets on FF++ and FakeAVCeleb.

Model	Offset	Backbone	Dataset	FF++	FakeAVCeleb
AVHuBERT [53]	Fixed	BASE	LRS3	95.3	97.0
		BASE	LRS3+Vox2	96.1	97.9
		LARGE	LRS3	95.7	96.8
		LARGE	LRS3+Vox2	**97.6**	99.0
	Dynamic	LARGE	LRS3+Vox2	93.7	98.7
VATLM [68]	Fixed	LARGE	LRS3+Vox2	97.1	**99.3**

Thanks!