Investigating self-supervised representations for audio-visual deepfake detection

Dragos-Alexandru Boldisor[12] Stefan Smeu[1] Dan Oneata[21] Elisabeta Oneata[1]

[1]Bitdefender, Bucharest [2]POLITEHNICA Bucharest

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation

- Prior work on deepfake detection uses self-supervised features in isolation or buried within complex architectures.
- Instead, we systematically evaluate a wide range of features and focus on three aspects: **performance**, **interpretability**, **complementarity**.

Models used

Model	Input Modality	Input Context	Pretraining Modality	Pretraining Domain
Wav2Vec2	A	full	A	speech
Auto-AVSR (ASR)	A	full	A	speech
AV-HuBERT (A)	A	full	AV	lips + speech
BRAVEn (A)	A	full	AV	lips + speech
CLIP VIT-L/14	V	frame	V	generic (images)
FSFM	V	frame	V	faces (video)
Video-MAE-large	V	chunk	V	generic (video)
Auto-AVSR (VSR)	V	full	V	lips (video)
AV-HuBERT (V)	V	full	AV	lips + speech
BRAVEn (V)	V	full	AV	lips + speech
Auto-AVSR	AV	full	AV	lips + speech
AV-HuBERT	AV	full	AV	lips + speech

Methodology: Linear probing

Given a video $\mathbf{x}$, we first **extract feature representations** $\phi(\mathbf{x})_t$ for each frame t, apply a **learnable linear classifier** $\mathbf{w}$, and **aggregate the predictions** using the log-sum-exp pooling function to get a **video-level prediction**:

$$s(\mathbf{x}; \mathbf{w}) = \log \sum_t \exp \left\{ \mathbf{w}^\intercal \phi(\mathbf{x})_t \right\}$$

By keeping the classifier minimal, we can assess the information encoded in the features. Skipping the final pooling function we obtain temporal explanations in the form of per-frame fakeness scores.

How useful are self-supervised features?

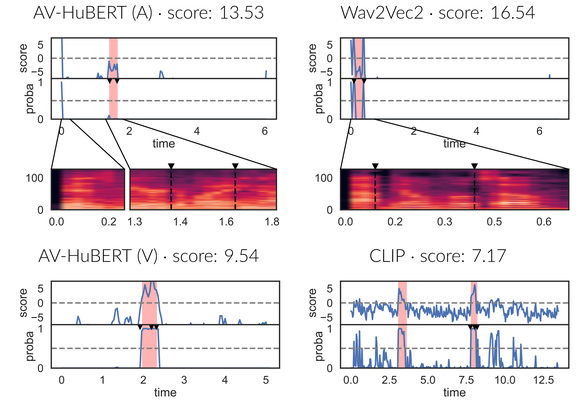
Model	Modality	ID FAVC	ID AV1M	ID DFE	OOD mean
AV-HuBERT (A) random	audio	**99.8**	**98.8**	53.4	**63.5**
AV-HuBERT (V) random	visual	83.8	50.2	**56.8**	55.1
Wav2Vec2	audio	**100**	**100**	58.7	**70.8**
Auto-AVSR (ASR)	audio	99.7	96.4	63.5	54.3
AV-HuBERT (A)	audio	**100**	**100**	65.8	67.3
BRAVEn (A)	audio	**100**	99.9	**67.8**	66.7
CLIP VIT-L/14	visual	99.8	96.5	73.9	63.2
FSFM	visual	97.1	95.3	71.7	55.0
Video-MAE-large	visual	**100**	**99.8**	54.5	55.6
Auto-AVSR (VSR)	visual	97.8	59.0	64.3	64.5
AV-HuBERT (V)	visual	**100**	93.7	72.1	80.0
BRAVEn (V)	visual	**100**	93.0	**75.5**	**84.6**
Auto-AVSR	audio-visual	94.7	91.6	61.2	54.7
AV-HuBERT	audio-visual	**100**	**99.9**	**70.4**	**78.2**

Performance of probed features: 1. Many features can encode deepfake information. 2. Audio-informed features generalize best. 3. Random features are not random. 4. In-the-wild DFE is harder.

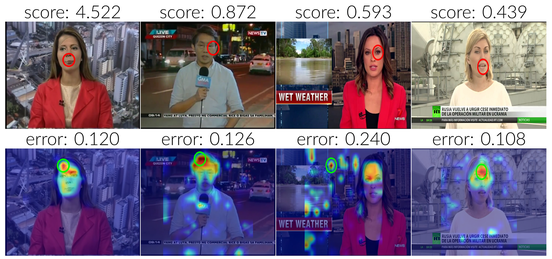
Method	Test dataset AV1M	FAVC	AVL	DFE	Average All4	Last3
Self-supervised features · Linear classifier trained on AV1M						
Wav2Vec2 (A)	**100**	**99.9**	56.3	58.6	78.7	71.6
BRAVEn (V)	93.0	98.8	**96.7**	**76.0**	**91.1**	**90.5**
AV-HuBERT (AV)	99.9	99.5	84.4	54.3	84.5	79.4
State-of-the-art approaches						
AVFF [5] · AV1M	98.0	89.2	53.9	58.1	74.8	67.1
AuViRe [3] · AV1M (all)	**100**	88.1	72.4	63.8	81.1	74.7
RealForensics [2] · FF++	60.7	89.1	76.8	74.4	75.2	80.1
AVAD [1] unsup.	52.9	95.2	73.2	64.8	71.5	77.7
SpeechForensics [4] unsup.	68.1	**100**	**92.7**	**75.6**	**84.1**	**89.4**

Comparison to SotA: Best set of features outperforms more complex models, despite using a simple linear classifier.

Where do self-supervised features look?

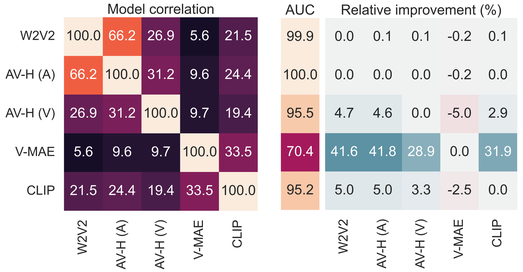


Temporal explanations (per-frame fakeness scores): Models focus on correct regions, but audio models also pick leading silence.



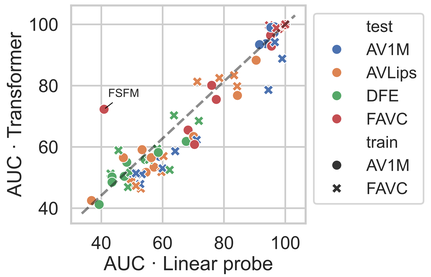
Spatial explanations (Grad-CAM on CLIP features): Model attends to facial artifacts and is aligned to human annotations (red circles).

How complementary are different features?



Correlation between model predictions: Vision models are more complementary than audio ones. **Relative improvement from late fusion**: Complementarity tends to help downstream performance.

How important is the classification head?



Performance of a Transformer head vs linear head: Features matter more than the classifier.

[1] Chao Feng et al. Self-supervised video forensics by audio-visual anomaly detection. In *Proc. CVPR*, 2023.

[2] Alexandros Haliassos et al. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proc. CVPR*, 2022.

[3] Christos Koutlis et al. AuViRe: Audio-visual speech representation reconstruction for deepfake temporal localization. In *Proc. WACV*, 2026.

[4] Yachao Liang et al. SpeechForensics: Audio-visual speech representation learning for face forgery detection. In *Proc. NIPS*, 2024.

[5] Trevine Oorloff et al. AVFF: Audio-visual feature fusion for video deepfake detection. In *Proc. CVPR*, 2024.