Multi-modal Deepfake Detection and Localization with FPN-Transformer

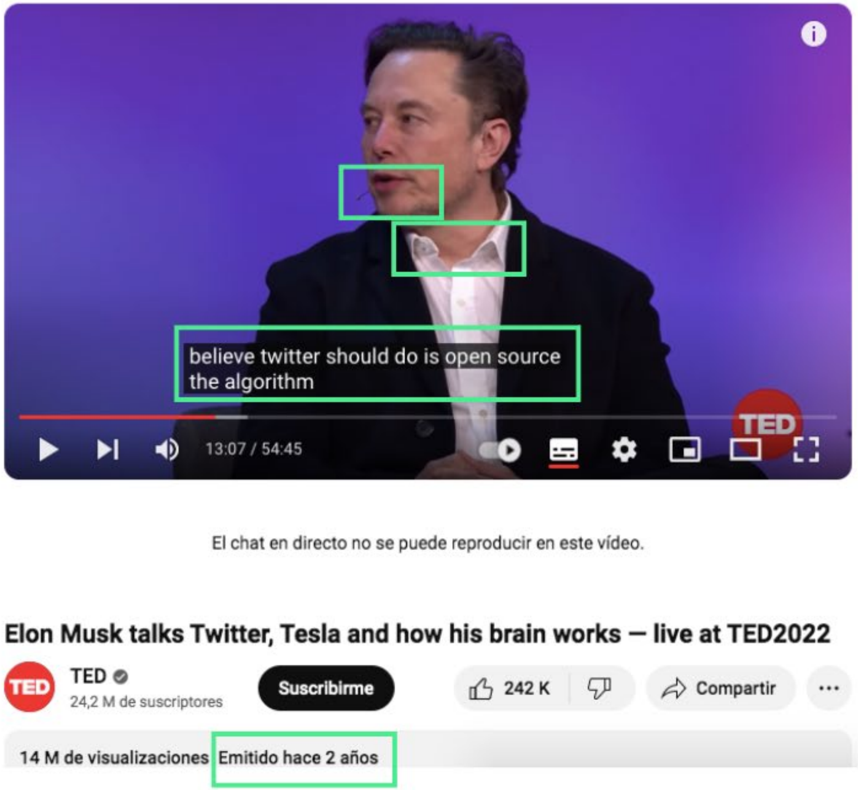
IJCAI 2025 Workshop
on Deepfake Detection, Localization, and Interpretability

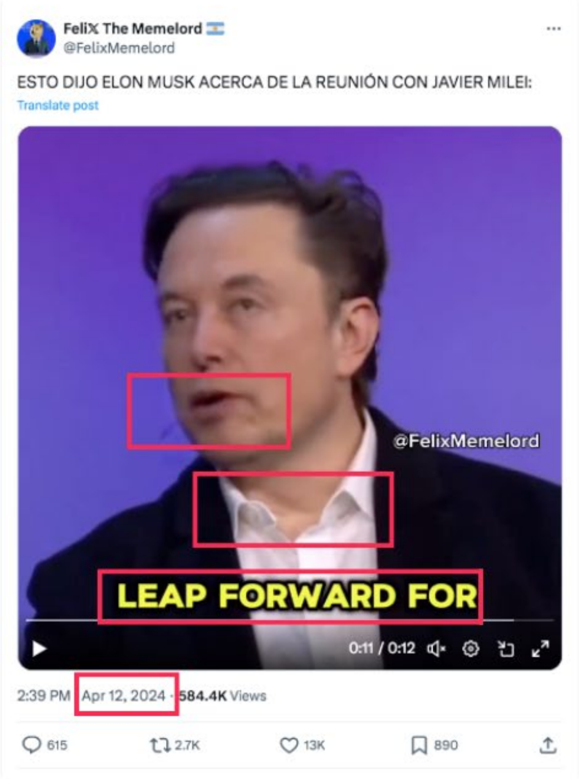
Speaker: Hezhe Qiao

Authors: Chende Zheng, Ruiqi Suo, Zhoulin Ji, Jingyi Deng, Fangbin Yi, Chenhao Lin, Chao Shen
Xi'an Jiaotong University

Backgrounds

Audio-Visual Deepfake Detection





Backgrounds

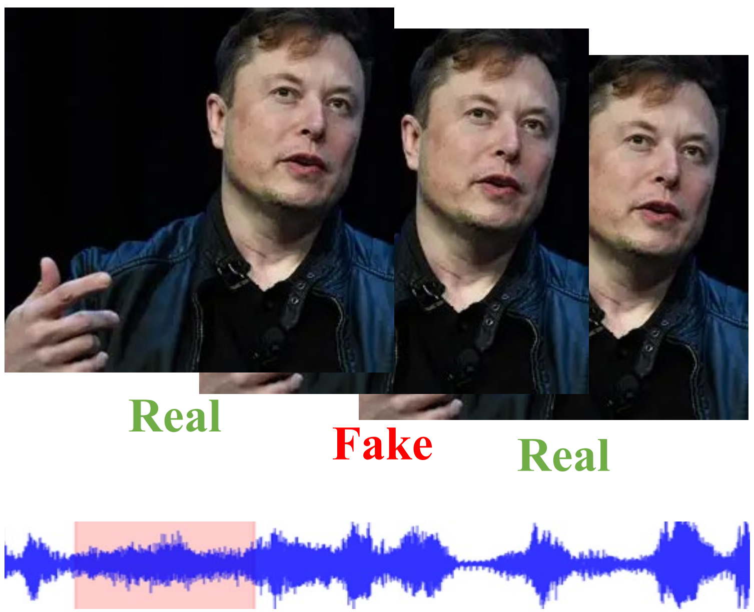
Main limitations

- *Multi-modality*
- *Precisely localization*
- *Diverse generators*

Manipulated Modality	Deepfake Methods Audio	Deepfake Methods Video	#Fake
V	0	4	4K
V	0	1	5K+
V	0	8	0.1M+
AV	1	3	0.2M+
V	0	8	0.1M+
A	3	0	0.5M+
AV	1	1	0.1M+
AV	2	1	0.8M+
AV	**9**	**18**	**0.3M+**

DDL-AV Datasets

Partial Forgery



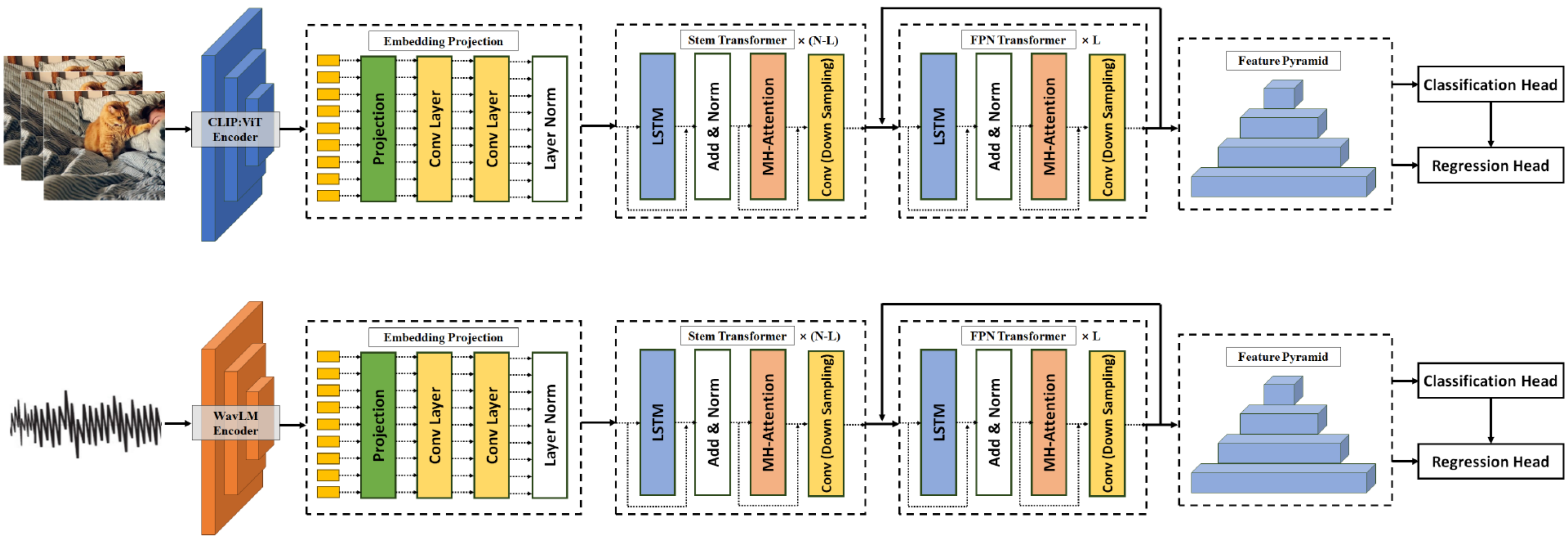
Detector:

Fake

But where?

Method - Overall

Dual FPN-Transformer Detection Framework



Three key components:

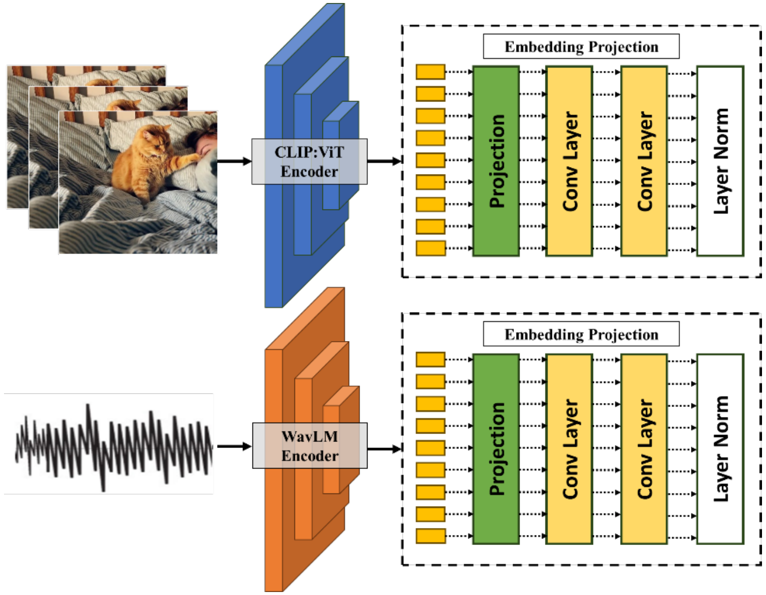
- Temporal feature embedding and projection module
- FPN-Transformer backbone module
- Classification and prediction heads

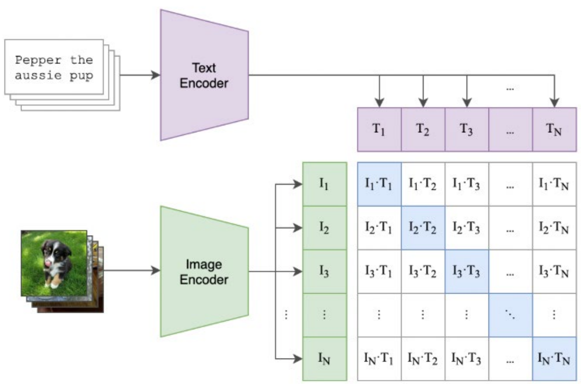
Feature Embedding

- Pre-trained **self-supervised models** are utilized as feature encoders.

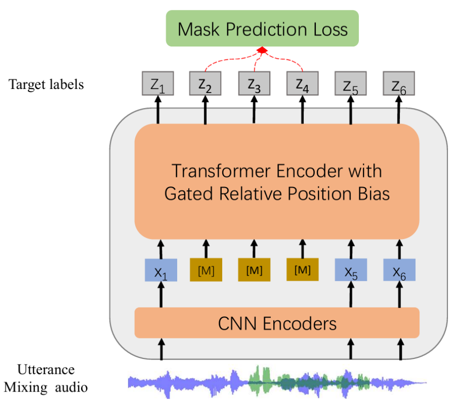
- We employ a set of **masked differential convolutional networks** to implement feature projection

$$\mathrm{MDC}(t_0) = \theta \cdot \left(-z(t_0) \cdot \sum_{t_n \in D} w(t_n) \right) + \sum_{t_n \in D} w(t_n) \cdot z(t_0 + t_n)$$





CLIP:ViT for video

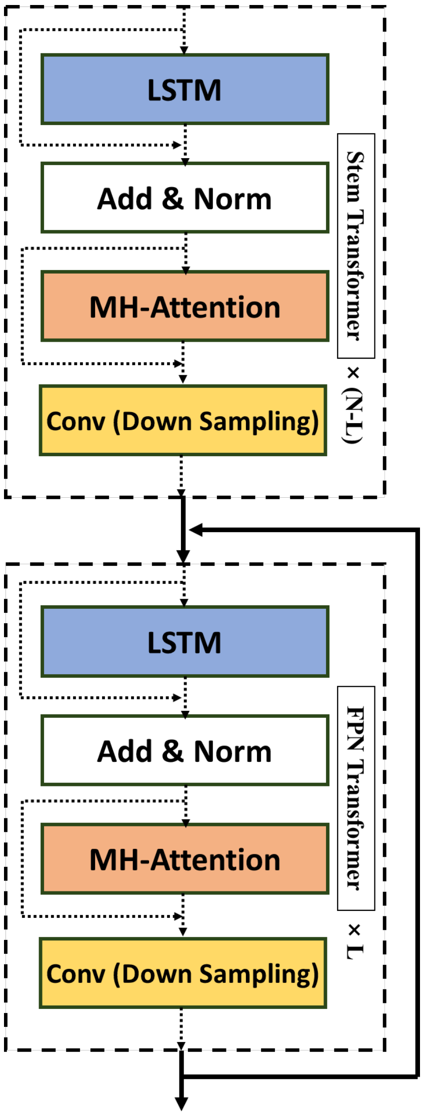


WavLM for audio

FPN-Transformer Architecture

- **We employ N layers of R-TLM blocks to perform deep feature encoding**
  - R-TLM incorporates **additional LSTM and Fusion layers** to explicitly model cross-context representation interactions.

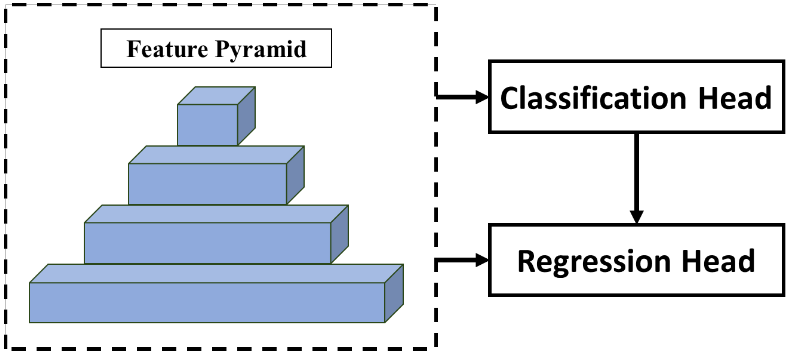
- **We introduce a strided depthwise 1D convolution after each MSA layer.**
  - By aggregating outputs from multi-level R-TLM structures, we obtain a hierarchical feature pyramid $F = \{F(1), \ldots, F(L)\}$ with L levels.



Dual-Branch Prediction

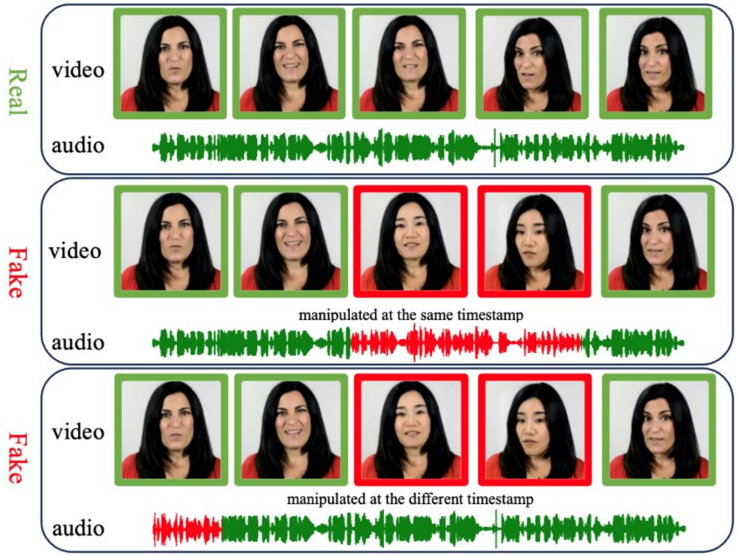
- **Classification Head**
  - We employ several 1D convolutional networks attached to each pyramid level, and the the classification head evaluates all L pyramid levels at **each timestamp t** to predict the **forgery probability $p(t)$.**

- **Regression Head**
  - The regression head predicts temporal boundaries only when timestamp t lies within forged segments
  - For each pyramid level, we predefine an output regression range to model the **start offset d_t^s** and **end offset d_t^e.**
  - The regression head employs 1D convolutional networks with ReLU activation to ensure precise distance estimation.



Experiment - Dataset

- **DDL-AV dataset**
  - 200k videos for training
  - 20k videos for validation
  - 111k videos for evaluation
  - 9 audio forgery methods
  - 18 video forgery methods



Examples of DDL-AV dataset

Datasets	Year	Tasks	Manipulated Modality	Deepfake Methods		#Fake
				Audio	Video	
FaceForensics++ [Rossler *et al.*, 2019]	2019	Cla	V	0	4	4K
Celeb-DF [Li *et al.*, 2020]	2020	Cla	V	0	1	5K+
DFDC [Dolhansky *et al.*, 2020]	2020	Cla	V	0	8	0.1M+
FakeAVCeleb [Khalid *et al.*, 2021]	2021	Cla	AV	1	3	0.2M+
ForgeryNet [He *et al.*, 2021]	2021	Cla/TL	V	0	8	0.1M+
ASVSpoof2021DF [Liu *et al.*, 2023]	2021	Cla	A	3	0	0.5M+
LAV-DF [Cai *et al.*, 2022]	2022	Cla/TL	AV	1	1	0.1M+
AV-Deepfake1M [Cai *et al.*, 2024]	2024	Cla/TL	AV	2	1	0.8M+
DDL-AV (ours)	**2025**	**Cla/TL**	**AV**	**9**	**18**	**0.3M+**

Experiment - Results

➢ **We compared the performance of different training strategy and self-supervised features.**

Training strategy		Feature Embedding		Final Score
Initial Learning Rate	Epochs	Audio	Video	
1×10^{-3}	3	wavLM	CLIP	0.7535
1×10^{-3}	6	wavLM	CLIP	0.7501
1×10^{-3}	15	wavLM	CLIP	0.6590
1×10^{-3}	36	wavLM	CLIP	0.6174
1×10^{-3}	60	wavLM	CLIP	0.6144
1×10^{-3}	95	wavLM	CLIP	0.6000
1×10^{-3}	6	wav2vec	XCLIP	0.7361
1×10^{-3}	60	wav2vec	XCLIP	0.5644
1×10^{-3}	95	wavLM	XCLIP	0.5873
1×10^{-3}	95	wav2vec	XCLIP	0.5798

1. WavLM + CLIP outperforms alternatives
2. Optimal training depth is critical
3. Framework robust to feature extractor variations

Experiment – Ablation on epochs

- **We compared the performance of different training epochs with a lower initial learning rate.**

Training strategy		Feature Embedding		Final Score
Initial Learning Rate	Epochs	Audio	Video	
3×10^{-4}	5	wavLM	CLIP	0.7164
3×10^{-4}	6	wavLM	CLIP	0.7218
3×10^{-4}	7	wavLM	CLIP	0.7218
3×10^{-4}	8	wavLM	CLIP	0.7340
3×10^{-4}	9	wavLM	CLIP	0.7252
3×10^{-4}	10	wavLM	CLIP	0.7201
3×10^{-4}	11	wavLM	CLIP	0.7256
3×10^{-4}	12	wavLM	CLIP	0.7227
3×10^{-4}	13	wavLM	CLIP	0.7182
3×10^{-4}	14	wavLM	CLIP	0.7126
3×10^{-4}	15	wavLM	CLIP	0.7084

1. The proposed method performs best at epoch 8
2. Deeper learning might lead to overfitting

Experiment – Visualization

➢ **Our proposed method could accurately detect and locate audio and video forgery.**

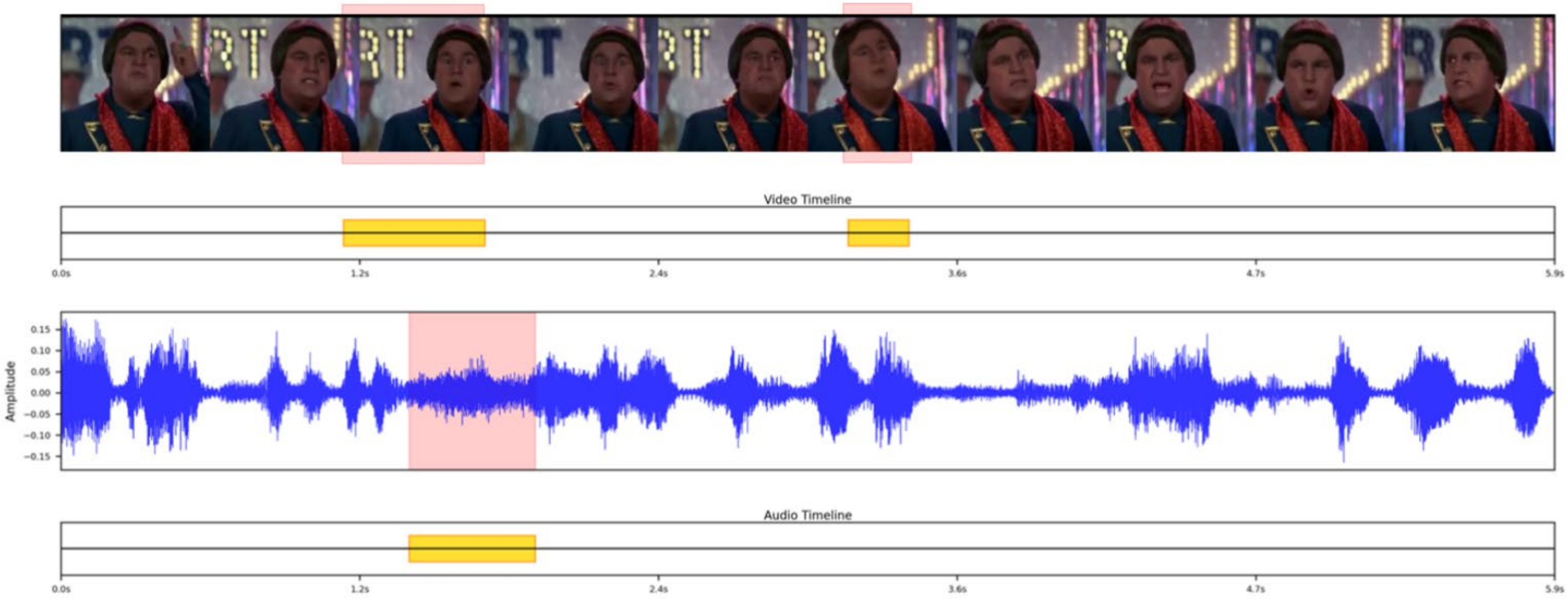


Figure 2: Visualization results of our proposed method. Red represents forged segments, and yellow represents our predicted results. Our method can accurately predict the presence of forged video and audio segments in the samples for both video and audio modalities.

Conclusion & Future Work

➢ Main Contributions

- ➢ A general-purpose temporal data forgery detection model for multimodal deepfake localization
- ➢ Extensive experiments on the IJCAI'25 DDL-AV dataset to validate the effectiveness

➢ Insights & Future work

- ➢ Leveraging pre-trained self-supervised models (WavLM for audio, CLIP for video) can effectively detect artifacts.
- ➢ Explore more modal information interaction for precise deepfake detection and localization.

Ending

Thank you for listening

Contact

Code