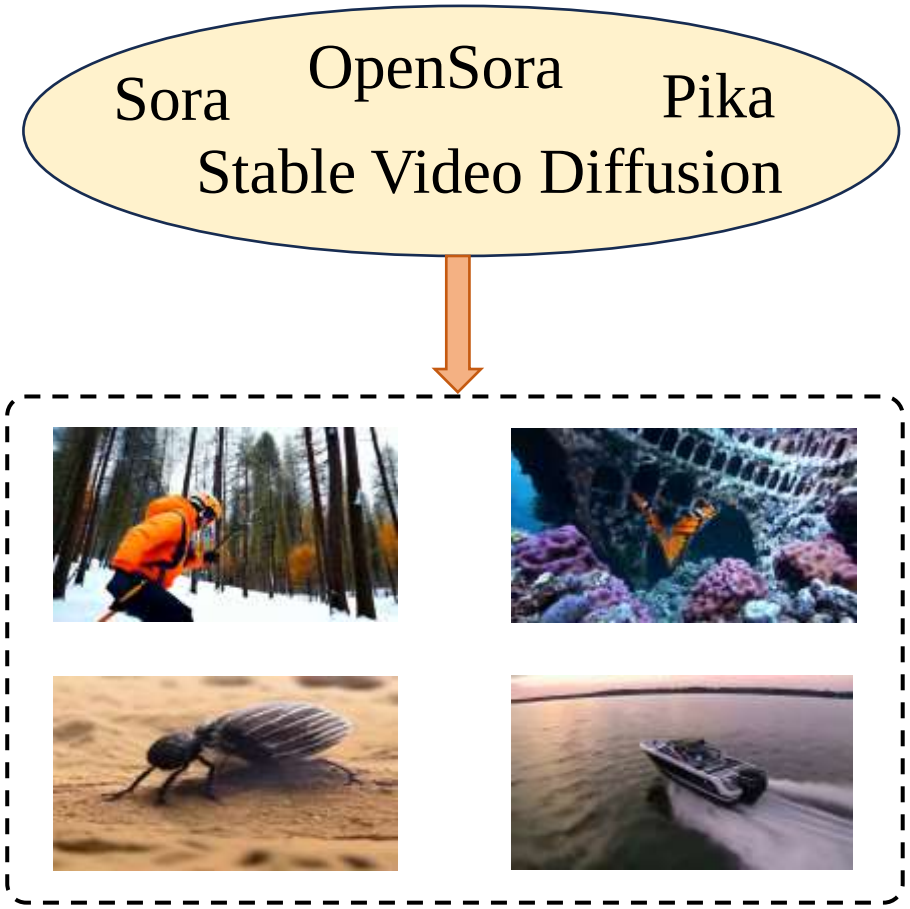
On Learning Multi-Modal Forgery Representation for Diffusion Generated Video Detection

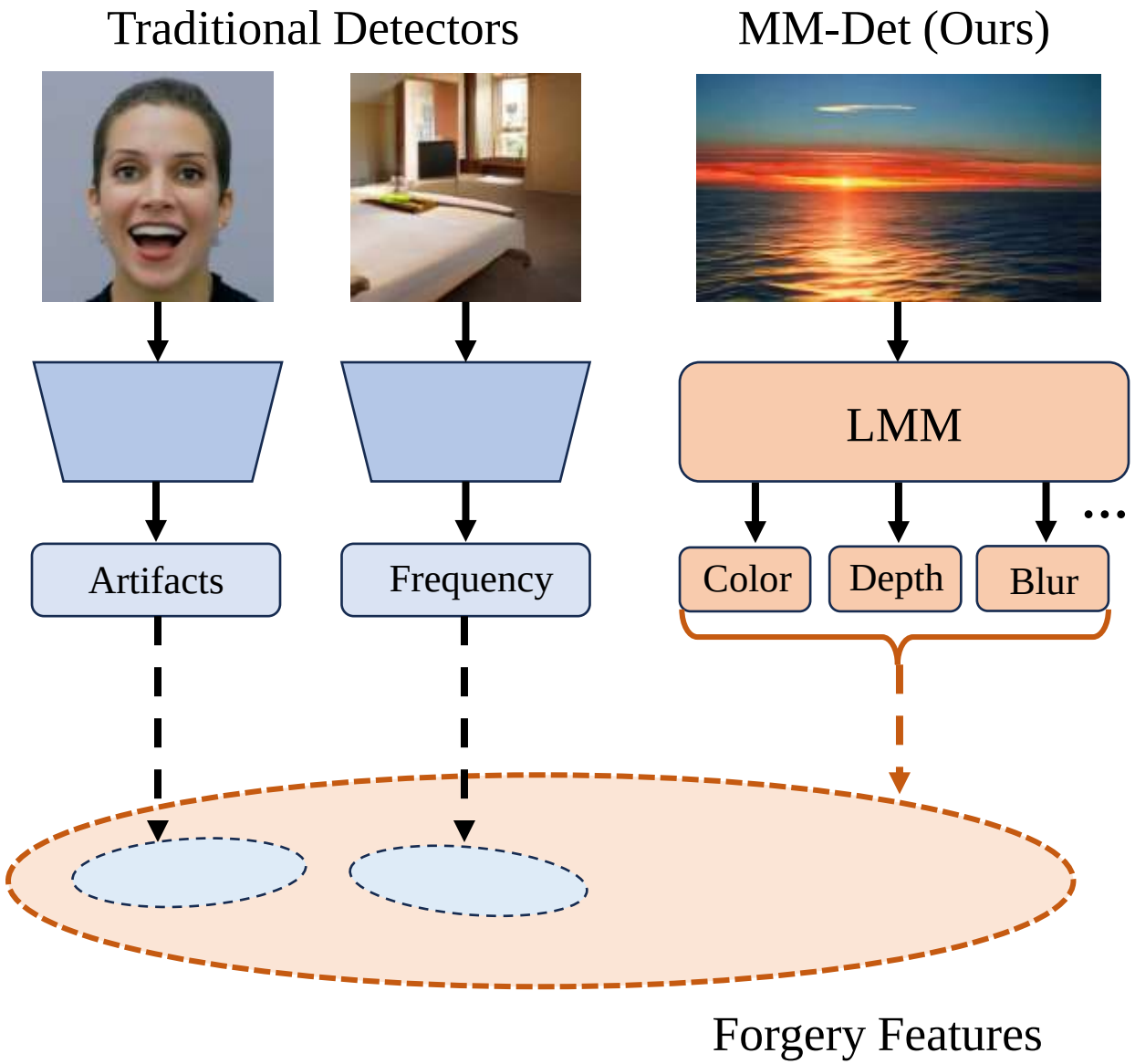
[1]Shanghai Jiao Tong University, [2]Michigan State University, [3]Shanghai Artificial Intelligence Laboratory

Xiufeng Song[1], Xiao Guo[2], Jiache Zhang[1], Qirui Li[1], Lei Bai[3], Xiaoming Liu[2], Guangtao Zhai[1], Xiaohong Liu[†1]

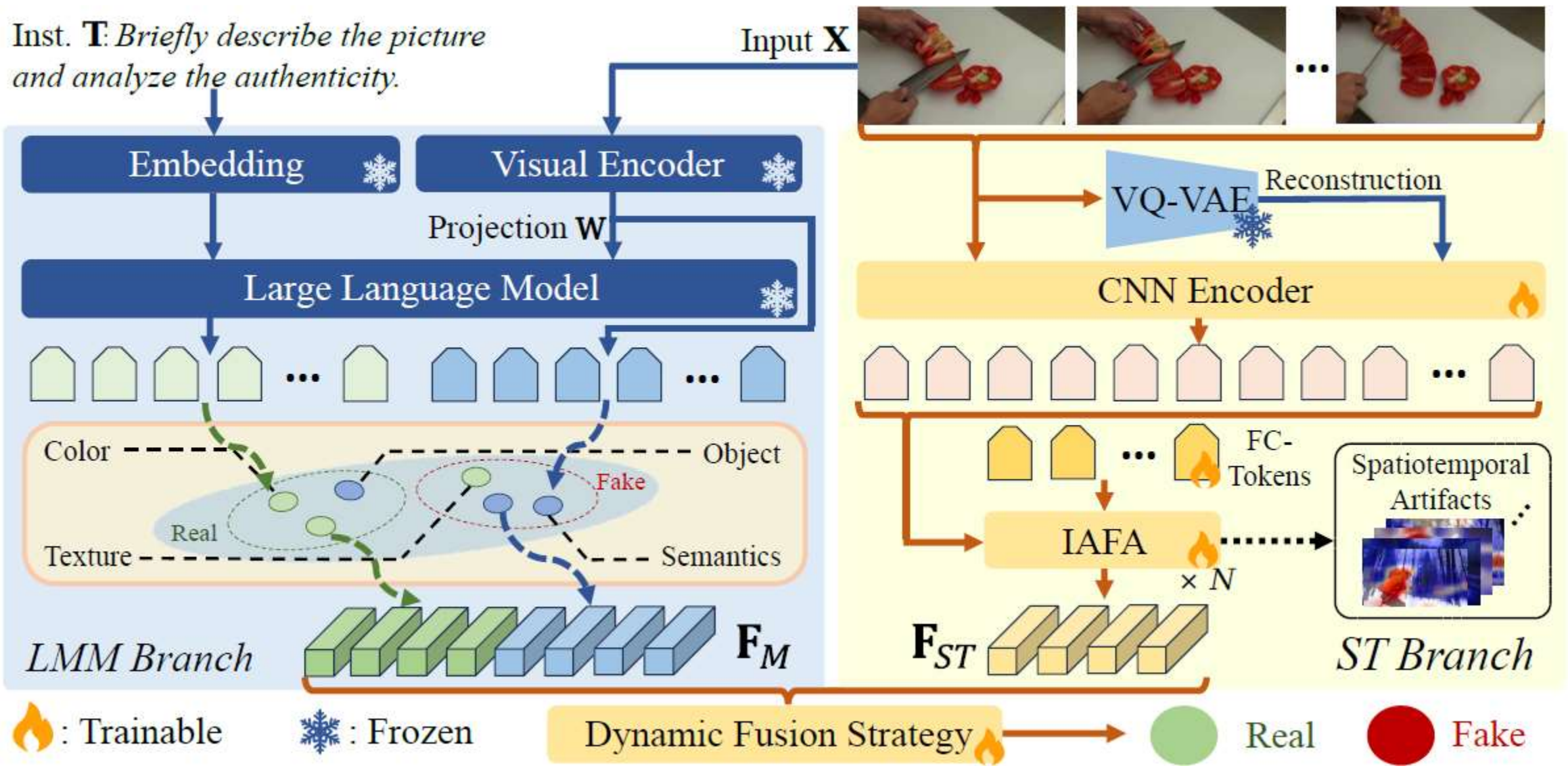
[†] Corresponding Author.

Introduction

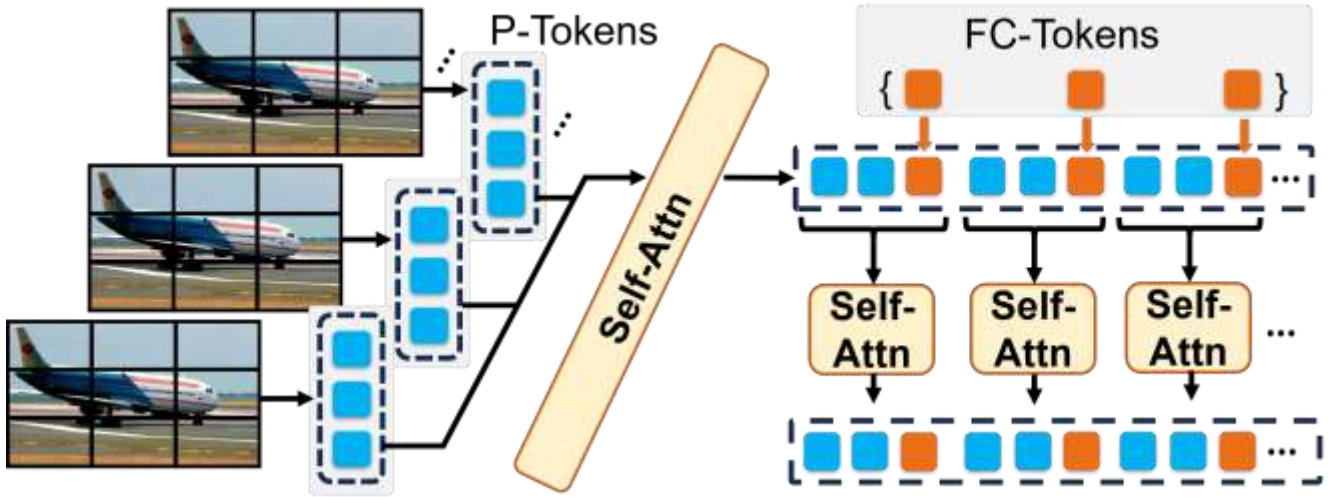




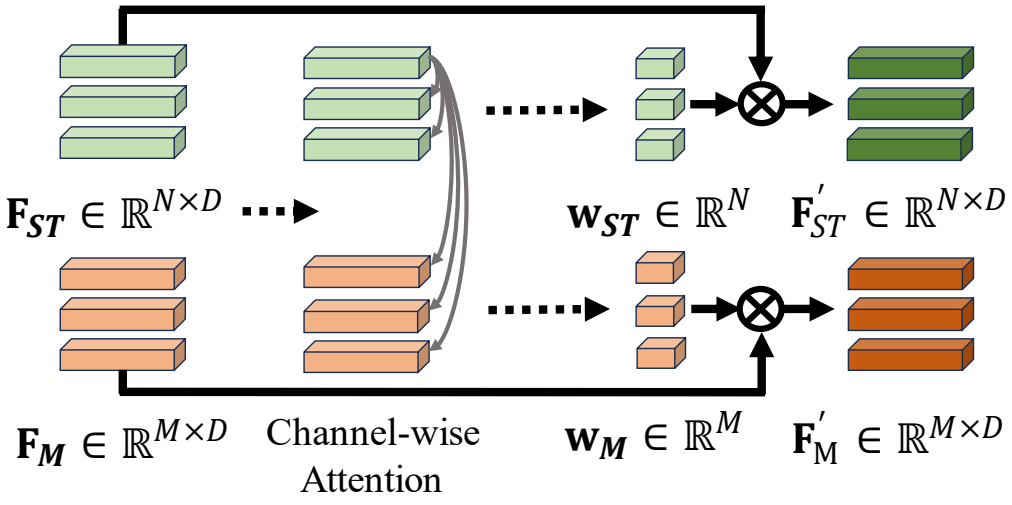
Overview of MM-Det



Method

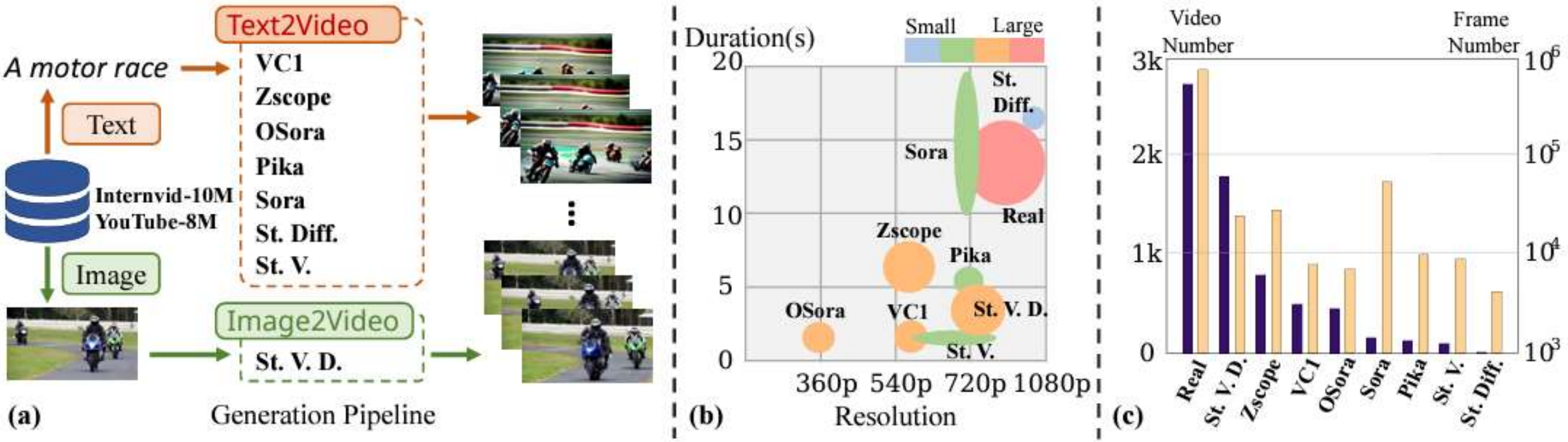


(a) Mechanism of In-And-Across Frame Attention (IAFA)



(b) Dynamic Fusion Strategy

Diffusion Video Forensics (DVF) Dataset

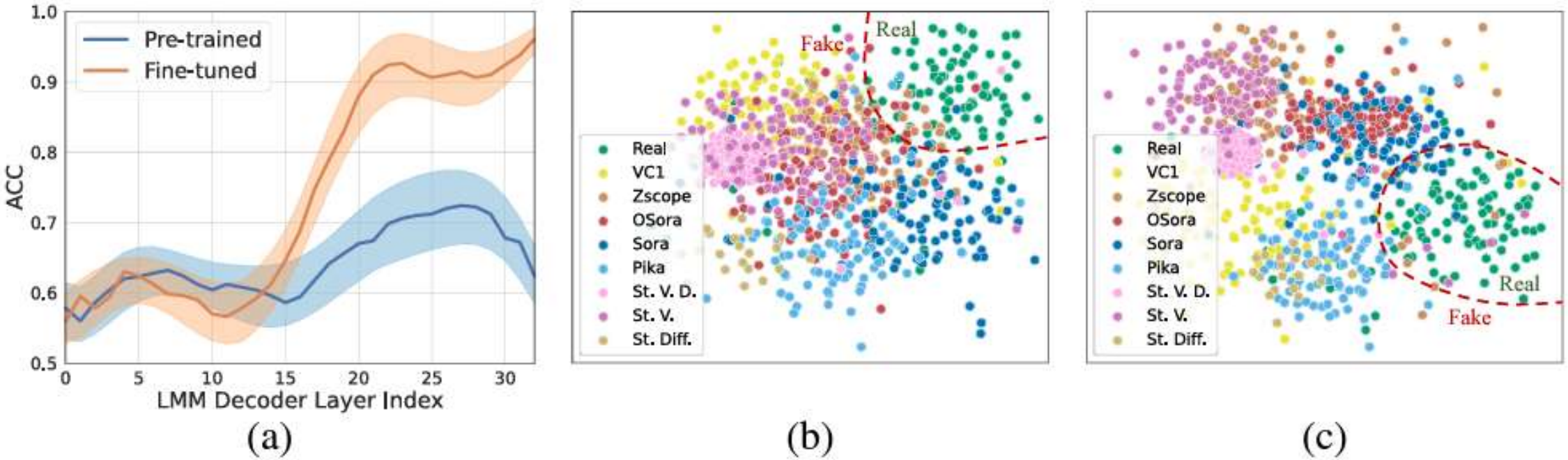


Performance

Table 1: Video forgery detection performance on the DVF dataset measured by AUC (%). [Key: **Best**; Second Best; Stable Diff.: Stable Diffusion; Avg.: Average]

Method	Video-Crafter1	Zero-scope	Open-Sora	Sora	Pika	Stable Diff.	Stable Video	Avg.
CNNDet [53]	83.3	70.2	81.9	63.8	76.5	71.8	80.8	75.5
DIRE [55]	56.8	61.9	56.1	60.7	70.0	58.3	71.2	62.1
Raising [8]	63.9	58.5	64.6	62.4	66.0	91.3	59.5	66.6
Uni-FD [36]	93.6	90.1	83.9	85.4	93.0	81.5	87.9	87.9
F3Net [38]	96.1	91.8	85.9	66.0	95.6	86.3	**96.0**	88.2
ViViT [3]	89.2	88.0	85.2	81.6	92.7	88.1	92.1	88.1
TALL [62]	76.5	61.8	69.8	62.3	79.9	85.9	64.8	71.6
TS2-Net [31]	60.7	72.0	74.3	81.0	80.2	60.2	80.2	72.7
DE-FAKE [43]	72.3	70.3	53.6	67.3	88.4	86.0	74.1	73.1
HiFi-Net [18]	96.7	93.9	94.9	83.9	85.8	80.2	87.3	89.0
MM-Det (Ours)	**97.4**	**98.6**	**97.6**	**91.7**	**98.0**	**92.1**	95.1	**95.7**

Analysis



(a) Clustering accuracy of features from different layers in LMM branch. (b)(c): t-SNE visualization of features from ST and LMM branches. (b)Features from the ST branch. (c): Features from the LMM branch.

Thank you for watching.