

DeformTrace: A Deformable State Space Model with Relay Tokens for Temporal Forgery Localization

Xiaodong Zhu, Suting Wang, Yuanming Zheng, Junqi Yang, Yangxu Liao, Yuhong Yang*, Weiping Tu, Zhongyuan Wang

National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, China
xiaodongzhu@whu.edu.cn, yangyuhong@whu.edu.cn, wzy_hope@163.com

Abstract

Temporal Forgery Localization (TFL) aims to precisely identify manipulated segments in video and audio, offering strong interpretability for security and forensics. While recent State Space Models (SSMs) show promise in precise temporal reasoning, their use in TFL is hindered by ambiguous boundaries, sparse forgeries, and limited long-range modeling. We propose DeformTrace, which enhances SSMs with deformable dynamics and relay mechanisms to address these challenges. Specifically, Deformable Self-SSM (DS-SSM) introduces dynamic receptive fields into SSMs for precise temporal localization. To further enhance its capacity for temporal reasoning and mitigate long-range decay, a Relay Token Mechanism is integrated into DS-SSM. Besides, Deformable Cross-SSM (DC-SSM) partitions the global state space into query-specific subspaces, reducing non-forgery information accumulation and boosting sensitivity to sparse forgeries. These components are integrated into a hybrid architecture that combines the global modeling of Transformers with the efficiency of SSMs. Extensive experiments show that DeformTrace achieves state-of-the-art performance with fewer parameters, faster inference, and stronger robustness.

Introduction

The rapid development of generative AI has greatly advanced multimedia creation but also raised security concerns due to the ease of producing realistic forgeries. Most existing work (Feng, Chen, and Owens 2023; Oorloff et al. 2024; Smeu et al. 2025) focuses on binary forgery detection, whereas Temporal Forgery Localization (TFL) provides finer interpretability by identifying manipulated segments.

Early TFL methods, such as BA-TFD (Cai et al. 2022) and BA-TFD+ (Cai et al. 2023), rely on CNNs or multi-scale transformers for frame-level classification, but suffer from low precision and slow inference. Later models like UMMAFormer (Zhang et al. 2023) and DiMoDif (Koutlis and Papadopoulos 2024) improve performance with pretrained extractors and feature pyramids, yet remain computationally heavy. Thus, developing a compact yet effective TFL architecture is imperative. TFL is conceptually related to temporal action detection (TAD), as both require precise boundary

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

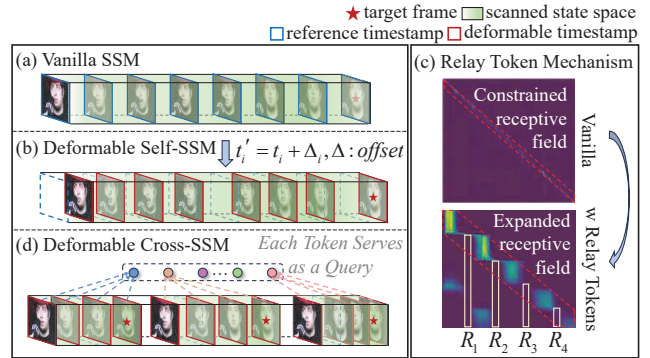


Figure 1: Overview of our main contributions (we take the video sequence for illustration). (a) Vanilla SSM; (b) Deformable Self-SSM with learnable temporal offsets for flexible local sampling; (c) Visualization of hidden attention, where relay tokens expand receptive fields and maintain long-range dependencies; (d) Deformable Cross-SSM enables cross-sequence interactions by allowing each query token to partition the global state space into subspaces.

localization in long sequences. Recent progress in TAD using state space models (SSMs) and query-based designs has provided valuable insights for model design.

Recent advances in State Space Models (SSMs), especially Mamba (Gu and Dao 2024; Dao and Gu 2024), demonstrate strong performance with lower complexity and faster inference. Models such as MambaIRv2 (Guo et al. 2025) and ActionMamba (Chen et al. 2024) outperform baselines in tasks like segmentation and temporal action detection, showing the potential of SSMs for dense, temporally-aware prediction. However, SSM-based architectures have not been applied to TFL due to three main challenges: **Boundary ambiguity**: Forgery boundaries are often unclear, unlike the well-defined ones in TAD. Standard SSMs use fixed state updates, causing temporal smoothing that reduces localization precision; **Sparse forgeries**: Most frames are non-forged, so SSMs’ recursive updates are dominated by non-forgery patterns, weakening sensitivity to sparse forgeries; **Limited long-sequence modeling**: Although efficient for long sequences, SSMs suffer from information decay over distance (Ye et al. 2025), limiting their

ability to capture long-range context.

To address boundary ambiguity, we propose Deformable Self-SSM (DS-SSM), which, for the first time, integrates a deformable dynamic receptive field mechanism into state space models. Unlike deformable Mamba variants in the image domain (Liu et al. 2025; Hu et al. 2024), our design leverages inherent temporal continuity of video and audio, thus omits operations such as patch splitting and token ranking. Instead, DS-SSM predicts offsets at each time step to dynamically sample input features, significantly reducing computational overhead while preserving SSMs’ low complexity. Figure 1(b) shows that, compared to vanilla SSM (see Figure 1(a)), DS-SSM better captures semantically relevant context beyond local fixed windows and improves robustness to ambiguous temporal boundaries.

Meanwhile, we introduce a Relay Token Mechanism into the DS-SSM to maintain long-range information flow, inspired by relay nodes in wireless communication. As shown in Figure 1(c), which compares the hidden attention of vanilla SSM and our variant with relay tokens (visualized following (Ali, Zimmerman, and Wolf 2025)), these periodically inserted learnable tokens effectively expands the receptive field and mitigates the long-range decay problem.

We further propose the Deformable Cross-SSM (DC-SSM), which introduces cross-sequence interactions into deformable state space modeling to tackle sparse forgeries. As shown in Figure 1(d), each auxiliary token, representing a potential forgery, serves as a query to retrieve forgery-relevant information from the main stream sequence. This mechanism partitions the global state space into query-specific subspaces, reducing non-forgery accumulation and improving sensitivity to sparse forgeries. Like cross-attention, DC-SSM enables explicit token-to-sequence interactions for targeted retrieval. Finally, These components are integrated into a hybrid TFL architecture combining Transformer attention with efficient state updates.

To sum up, the main contributions of this work are: (1) We propose Deformable Self-SSM (DS-SSM), the first to introduce dynamic receptive fields into temporal state space models, improving localization of ambiguous boundaries. (2) We introduce a Relay Token mechanism that explicitly mitigates the long-range decay of SSMs, which is a key limitation in prior state-space models. (3) We propose Deformable Cross-SSM (DC-SSM), the first to incorporate cross-sequence interactions into state space modeling, enhancing sensitivity to sparse forgeries. (4) We integrate these components into a unified hybrid TFL architecture that combines the strengths of Transformers and SSMs. Extensive experiments on standard benchmarks demonstrate state-of-the-art performance with lower model size and faster inference.

Related Work

State Space Model Applications

State Space Models (SSMs) provide an efficient approach to sequence modeling. Mamba (Gu and Dao 2024; Dao and Gu 2024) builds on this by introducing a selective mechanism that enables long-sequence modeling with linear complexity. Mamba has inspired a growing number of variants in vision

tasks, focusing on scalability, efficiency, and local-global modeling (Zhu et al. 2024; Liu et al. 2024b; Pei, Huang, and Xu 2025; Patro and Agneeswaran 2024; Hatamizadeh and Kautz 2025). Meanwhile, deformable variants (Liu et al. 2025; Hu et al. 2024) incorporate deformable scanning to handle irregular patterns but they are restricted to uni-sequence modeling. Moreover, both Mamba and its variants suffer from long-range decay. LongMamba (Ye et al. 2025) attempts to solve this using channel classification and token filtering, but its reliance on fixed thresholds limits adaptability.

Temporal Action Detection

Temporal Action Detection (TAD) aims to detect and localize actions in untrimmed videos. Early anchor-based methods (Bai et al. 2020; Lin et al. 2018, 2019; Long et al. 2019) struggle with varying action lengths due to fixed anchors. Anchor-free methods (Cheng and Bertasius 2022; Lin et al. 2021; Zhang, Wu, and Li 2022) improve flexibility via center-based or asymmetric designs. Query-based approaches like TadTR (Liu et al. 2022) and TE-TAD (Kim et al. 2024), inspired by DETR (Carion et al. 2020), treat TAL as set prediction, removing anchors and post-processing (Tan et al. 2021). Recently, Mamba-based architectures (Chen et al. 2024; Liu et al. 2024a; Wang et al. 2025) have been introduced to TAD, enhancing temporal modeling.

Temporal Forgery Localization

Temporal Forgery Localization (TFL) aims to identify forged segments in videos, requiring fine-grained temporal modeling. While most research focuses on video-level detection (Zhou and Lim 2021; Yang et al. 2023; Yu et al. 2023), TFL remains underexplored. Recent datasets such as Lav-DF (Cai et al. 2023) and AV-Deepfake1M (Cai et al. 2024) have started to drive progress in this direction. BA-TFD (Cai et al. 2022) introduced contrastive loss for frame-level classification, followed by BA-TFD+ (Cai et al. 2023) with multi-scale Transformers. UMMAFormer (Zhang et al. 2023), DiMoDif (Koutlis and Papadopoulos 2024), MFMS (Zhang et al. 2024), and CLFormer (Cheng et al. 2025) improved accuracy via attention and feature pyramids. Vigo (Pérez-Vieites et al. 2024) employs an IoU-based fusion strategy, while TransHFC (Huang et al. 2025) integrates hypergraph learning and Transformers. Zhu et al. (Zhu et al. 2025) proposed a query-based audio-visual framework for efficient end-to-end temporal localization.

The Proposed Method

Overall Architecture

Given a dataset of untrimmed talking face videos, each sample $x = \{x^v, x^a\}$ contains temporally aligned video and audio streams. However, parts of x may be manipulated in either modality. The goal of DeformTrace is to localize forgery segments $\mathcal{A} = \{(t_i, d_i)\}_{i=1}^{N_f}$, where t_i and d_i represent the center timestamp and duration of the i -th forgery, respectively. Here, N_f denotes the total number of forgeries. Additionally, it predicts a binary authenticity label y .

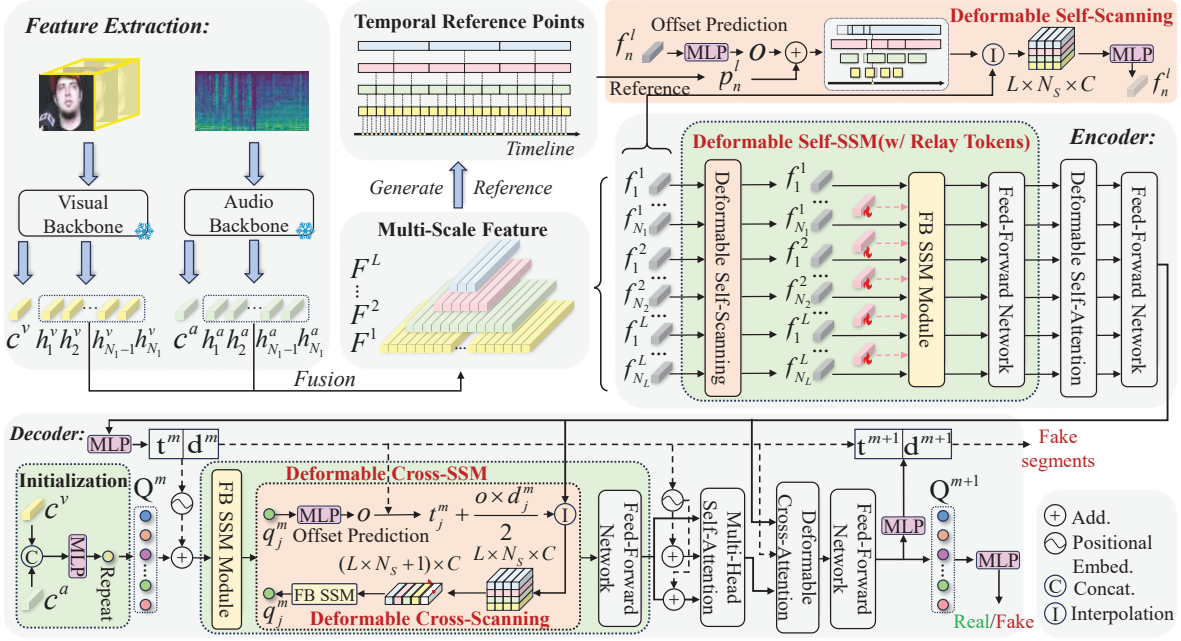


Figure 2: Illustration of the overall scheme of DeformTrace. Built on TadTR (Liu et al. 2022), DeformTrace integrates a multi-scale audio-visual feature extraction module, a deformable encoder for temporal modeling, and a deformable decoder for forgery localization and video-level classification. By incorporating deformable self- and cross-SSM modules, it combines Mamba’s efficient state updates with Transformer’s global modeling. Relay tokens with enhanced and cooperation losses help preserve long-range information dependencies during self-scanning.

Feature Extraction. As illustrated in Figure 2, DeformTrace adopts a query-based architecture based on TadTR (Liu et al. 2022). The video $x^v \in \mathbb{R}^{T \times H \times W \times C_0}$ and the mel spectrogram $x^a \in \mathbb{R}^{T \times N_{\text{mel}}}$ are processed by frozen, pretrained visual and audio backbones to extract classification tokens $\{c^v, c^a\}$ and feature sequences $\{h_n^v, h_n^a\}_{n=1}^{N_1}$, where both the classification and feature tokens have a dimensionality of C . Here, T and N_1 represent the number of input frames and extracted tokens of the first scale. To preserve temporal resolution for fine-grained localization, we set $T = N_1$. These sequences are concatenated and linearly projected to form fused features F^1 , further processed by $(L-1) 2 \times$ downsampling layers to obtain features $\mathcal{F} = \{F^l\}_{l=1}^L$, capturing multi-scale temporal context.

Encoder. The encoder receives flattened features $\{f_n^l\}_{n=1}^{N_l}$, where n denotes the feature index of each scale and $N_l = N_1/2^{l-1}, l = 1, 2, \dots, L$, then the encoder outputs a context-enhanced sequence X_E . As shown in Figure 2, it comprises a Deformable Self-SSM, a Deformable Self-Attention (DSA) (Liu et al. 2022), and a Feed-Forward Network (FFN). The DSA adaptively attends to sparse temporal locations around a reference point, while the FFN adds nonlinearity and further refines features.

Decoder. Visual and audio classification tokens are fused via an MLP into a global multimodal feature, which is repeated N_q times to form initial queries $Q^0 = \{q_j^0\}_{j=1}^{N_q}$, providing coarse alignment with ground-truth and aiding convergence (Kim et al. 2024). Initial proposals $\mathcal{A}^0 =$

$\{(t_j^0, d_j^0)\}_{j=1}^{N_q}$, representing center and duration of segments, are generated from fused features via MLP. The decoder has M layers, each containing Deformable Cross-SSM, Multi-Head Self-Attention (MHSA), Deformable Cross-Attention (DCA) (Liu et al. 2022), and FFN. MHSA models inter-query relations using anchor-based positional embeddings, while DCA allows each query to attend to relevant audio-visual context. The DCA output is passed through the FFN to update queries and then through an MLP to predict offsets $\Delta t, \Delta d$ for proposal refinement. The final proposals \mathcal{A}^M are used as predicted forgery segments, and the final queries Q^M are passed through an MLP classifier to produce the video-level forgery score \hat{y} .

Deformable Self-SSM

The Deformable Self-SSM (DS-SSM) module comprises three components: Deformable Self-Scanning, Forward-Backward SSM (FB SSM) (Wang et al. 2023), and a Feed-Forward Network (FFN). Deformable Self-Scanning enables each token to flexibly sample semantically relevant positions beyond local windows. FB SSM performs bidirectional state-space propagation on the sampled sequence, capturing contextual information from both past and future.

For the n -th feature in the l -th layer of the multi-scale feature \mathcal{F} , denoted f_n^l , we compute a normalized reference location p_n^l to align with the temporal axis of the video:

$$p_n^l = \frac{\omega \cdot 2^{l-1} \cdot (n + 0.5)}{\text{fps} \cdot d}. \quad (1)$$

where fps denotes frames per second (default 25), ω is the stride that indicates how many frames to step when extracting features, and d is the video duration in seconds. The numerator above estimates the temporal center of f_n^l in terms of frame index, converts it to seconds by dividing by fps, and normalizes the result by d , yielding a value in $(0, 1]$. We pre-compute these temporal reference points to accelerate the sampling process.

Given a feature f_n^l and its temporal reference point p_n^l , we predict an offset matrix $o \in \mathbb{R}^{L \times N_s}$ via a multi-layer perceptron (MLP), where N_s is the number of offsets per scale. For each feature, multiple offsets are predicted across scales, enhancing the flexibility and diversity of sampling. p_n^l is repeated and added element-wise with o to obtain deformable points \hat{p} . A bilinear interpolation function $\phi(\cdot, \cdot)$ extracts features at positions \hat{p} from multi-scale feature $\{F^l\}_{l=1}^L$, which are then aggregated via another MLP to produce the output feature \hat{f}_n^l . The process is formulated as:

$$\begin{aligned} o &= \text{MLP}_1(f_n^l), \\ \hat{p} &= \text{Repeat}(p_n^l) + o, \\ \hat{f}_n^l &= \text{MLP}_2(\phi(\hat{p}, \{F^l\}_{l=1}^L)). \end{aligned} \quad (2)$$

Relay Token Mechanism

Long-range decay is an inherent limitation of SSMs. Token interactions rely on the control matrix power \bar{A}^k , where k is the pairwise distance (Gu and Dao 2024; Dao and Gu 2024). Since \bar{A} contains values less than 1, interactions decay exponentially with distance. For instance, in DS-SSM, semantically aligned tokens f_1^1 and f_1^L are separated by $k \approx 2N_1$ (with $N_1 > 100$ in datasets like AV-Deepfake1M and LAV-DF), resulting in negligible interaction and limited receptive fields (see Figure 1(d)).

To address this, we draw inspiration from communication systems, where relay nodes mitigate long-distance signal attenuation. We propose a Relay Token Mechanism: learnable global tokens $\{r_i\}_{i=1}^{N_r}$ are evenly inserted into the input sequence before state-space updates. These input-independent tokens partition the sequence into $N_r + 1$ subspaces. Within each subspace, local states efficiently pass information to the relay token, which then broadcasts the aggregated message to other subspaces. This forms a sparse sequence-to-token information flow across the sequence. Figure 1(d) shows attention maps illustrating receptive field expansion and information relay via the relay tokens.

Deformable Cross-SSM

The Deformable Cross-SSM (DC-SSM) module comprises three components. The FB-SSM captures temporal dependencies around each query, while Deformable Cross-Scanning allows each query to dynamically attend to semantically relevant audio-visual features from the encoder, thereby enhancing the localization of forged regions. An FFN further refines the resulting features.

In Deformable Cross-Scanning, for the m -th layer of the decoder, each query $q_j^m \in Q^m$ uses its anchor proposal

(t_j^m, d_j^m) as a reference. A MLP predicts a multi-scale offset $o \in \mathbb{R}^{L \times N_s}$, and the deformable points are computed as $p_j^m = t_j^m + o \times d_j^m / 2$. A bilinear interpolation function samples features at p_j^m from the encoder output X_E , producing a feature tensor of shape $L \times N_s \times C$. This tensor is flattened and concatenated with a learnable empty token, forming a sequence of length $(L \times N_s + 1) \times C$. The sequence is then passed through a forward state space update. Owing to the aggregation behavior of SSMs, only the final token, the appended one, is retained as the output feature \hat{q}_j^m .

Losses and Optimization

In Relay Token Mechanism, we introduce two auxiliary losses: enhance loss and cooperation loss. The former encourages each relay token to better aggregate information from its neighboring sequence segments, while the latter promotes effective collaboration among relay tokens.

Enhance loss. For the k -th relay token r_k , we compute the average representation f_k^{avg} of its neighboring non-relay subsequences (excluding other relay tokens). The enhance loss is then defined as:

$$\mathcal{L}_{\text{enh}} = -\frac{1}{N_r} \sum_{k=1}^{N_r} \cos(r_k, f_k^{avg}), \quad (3)$$

Cooperation loss. To promote diversity and reduce redundancy among relay tokens, we introduce the *Cooperation Loss*, which minimizes the mutual information (MI) between different relay tokens. This can be approximated by encouraging the similarity matrix G of relay tokens to approach a scaled identity matrix:

$$\begin{aligned} \mathcal{L}_{\text{coop}} &= \min \sum_{i < j} I(r_i; r_j), \\ &\approx \|G - \gamma I\|^2 = \sum_{i \neq j} G_{ij}^2 + \sum_{k=1}^{N_r} (G_{kk} - \gamma)^2, \end{aligned} \quad (4)$$

where γ is a hyperparameter controlling the target self-similarity, we set it to 1.

Task-aware loss. Following DETR (Carion et al. 2020), we adopt the standard Hungarian Matching loss $\mathcal{L}_{\text{match}}$, which includes classification loss \mathcal{L}_{ce} and regression loss \mathcal{L}_{reg} . It is defined as:

$$\mathcal{L}_{\text{match}}(\mathcal{A}^M, \mathcal{A}) = \sum_{i=1}^{N_f} \sum_{u \in \{\text{ce}, \text{reg}\}} \mathcal{L}_u(\mathcal{A}_{\pi(i)}^M, \mathcal{A}_i), \quad (5)$$

where π denotes the optimal permutation from Hungarian Matching, $\mathcal{A}^M, \mathcal{A}$ denote the predicted and ground-truth segments, respectively. Besides, we apply a video-level cross-entropy loss between the predicted label \hat{y} and ground truth y , denoted as \mathcal{L}_{cls} .

Overall, the training loss is a weighted combination of four components:

$$\mathcal{L} = \mathcal{L}_{\text{match}} + \mathcal{L}_{\text{cls}} + \lambda_1 \cdot \mathcal{L}_{\text{enh}} + \lambda_2 \cdot \mathcal{L}_{\text{coop}}. \quad (6)$$

where λ_1 and λ_2 are hyperparameters controlling the contributions of the enhance and cooperation losses, respectively.

Type	Method	Venue	Modality	mAP(%)				mAR(%)				
				0.5	0.75	0.95	Avg.	100	50	20	10	Avg.
Anchor-based	BMN	ICCV'19	\mathcal{V}	24.0	7.60	0.10	10.6	53.3	41.2	31.6	26.9	38.3
	BMN(I3D)	ICCV'19	\mathcal{V}	10.6	1.70	0.00	4.10	48.5	44.4	37.1	31.6	40.4
Anchor-free	AGT	Arxiv'21	\mathcal{V}	17.9	9.40	0.10	9.10	43.2	34.2	24.6	16.7	29.7
	ActionFormer	ECCV'22	\mathcal{V}	95.3	90.2	23.7	69.7	88.4	89.6	90.3	90.4	89.7
	TriDet	CVPR'23	\mathcal{V}	96.3	86.8	23.6	68.9	91.0	90.4	90.0	88.7	90.0
	MDS	MM'20	\mathcal{AV}	12.8	1.60	0.00	4.80	37.9	36.7	34.4	32.2	35.3
	AVFusion	Arxiv'21	\mathcal{AV}	65.4	23.9	0.10	29.8	63.0	59.3	54.8	52.1	57.3
	BA-TFD	DICTA'22	\mathcal{AV}	76.9	38.5	0.30	38.6	66.9	64.1	60.8	58.4	62.5
	BA-TFD+	CVIU'23	\mathcal{AV}	96.3	85.0	4.40	61.9	81.6	80.5	79.4	78.8	79.8
	UMMAFormer	MM'23	\mathcal{AV}	98.8	95.5	<u>37.6</u>	77.3	92.4	92.5	92.5	<u>92.1</u>	<u>92.3</u>
Query-based	ELF-MDC	TCSVT'24	\mathcal{AV}	94.9	74.9	1.90	57.2	76.1	74.2	72.2	71.2	73.4
	DiMoDif	Arxiv'24	\mathcal{AV}	95.5	87.9	20.6	67.8	94.2	93.7	<u>92.7</u>	91.4	91.9
	TadTR	TIP'22	\mathcal{V}	80.2	61.1	5.20	48.8	72.9	72.5	70.56	69.2	71.2
	TE-TAD	CVPR'24	\mathcal{V}	85.7	64.9	7.13	52.6	79.2	78.6	78.4	76.2	77.9
	FullFormer (Ours)	–	\mathcal{AV}	94.6	85.7	29.4	69.9	88.4	88.4	86.9	85.5	87.3
	DeformTrace (Ours)	–	\mathcal{AV}	<u>97.1</u>	<u>90.7</u>	38.1	<u>75.3</u>	<u>93.3</u>	<u>93.1</u>	92.8	92.3	92.9

Table 1: Temporal forgery localization results on LAV-DF (Cai et al. 2023) benchmark. The modality denotes the input type: \mathcal{V} for visual, \mathcal{A} for audio and \mathcal{AV} for both. The best results are **bolded**, and second-best underlined.

Experiments

Experiment Setup

Datasets and Evaluation Metrics. We conduct experiments on two audio-visual deepfake datasets: LAV-DF (Cai et al. 2023) and AV-Deepfake1M (Cai et al. 2024). LAV-DF contains 78K/31K/26K train/val/test samples. AV-Deepfake1M is a larger and more challenging benchmark with 746K/57K/343K samples and finer-grained forgeries, posing greater difficulty for TFL models. Following standard practice, we report mean Average Precision (mAP) at IoU thresholds $\{0.5, 0.75, 0.9, 0.95\}$, and mean Average Recall (mAR) at different numbers of proposals $\{5, 10, 20, 30, 50, 100\}$. Additionally, we evaluate video-level detection on AV-Deepfake1M using Area Under the Curve (AUC) as the metric. We compare model parameters and computational cost on a single RTX 3090 GPU, evaluating all methods on the full LAV-DF test set (avg. video length: 8.6 s) with total/trainable parameters, FLOPs, and inference time.

Implementation Details. We adopt Raven’s visual and audio encoders (Haliassos et al. 2023) for feature extraction, due to their strong performance on diverse downstream tasks. Specifically, we use the base model pretrained on VoxCeleb2 (Chung, Nagrani, and Zisserman 2018) and LRS3 (Afouras, Chung, and Zisserman 2018). Videos are sampled at 25 fps and audio at 16 kHz. Visual frames are resized to 224×224 pixels, and audio is converted into 80-bin mel-spectrograms using a 40 ms Hamming window with a 40 ms hop size. Each input spans 8 seconds, resulting in $T = N_1 = 200$ frames and $C = 256$ dimensions after encoding. We adopt $L = 6$ feature levels via 1D convolution with kernel size 3, and use $M = 3$ decoder layers. In each deformable self-/cross-SSM layer, we sample $N = 6$ points per scale. We set the number of query and relay tokens to $N_q = 60$ and $N_r = 8$, respectively. The total loss combines

two terms weighted by $\lambda_1 = 0.5$ and $\lambda_2 = 0.2$. We train for 30 epochs on AV-Deepfake1M and 100 epochs on LAV-DF with a batch size of 32, using AdamW (Loshchilov and Hutter 2019) and a cosine scheduler with 5-epoch warm-up. The learning rate is set to $2e-4$. All experiments are conducted on eight NVIDIA RTX 3090 GPUs.

Comparison Methods. For the temporal forgery localization task, our baseline systems include three publicly available audio-visual TFL methods: BA-TFD (Cai et al. 2022), BA-TFD+ (Cai et al. 2023), and UMMAFormer (Zhang et al. 2023). In addition, we construct a baseline by replacing all SSMs in DeformTrace with Transformer blocks, referred to as FullFormer, for comparison. To enable a comprehensive comparison, we also report results on the LAV-DF and AV-Deepfake1M benchmarks.

Comparison with State-of-the-arts

TFL results. Tables 1 and 2 compare our method with state-of-the-art TFL approaches on LAV-DF and AV-Deepfake1M. On both datasets, our method outperforms the pure Transformer baseline by over 7% in mAP and mAR, highlighting the effectiveness of SSMs and our proposed modules. On LAV-DF, it achieves top results on mAP@0.95 and mAP@ $\{20, 10, \text{Avg.}\}$, and ranks second on the remaining metrics, demonstrating robustness across evaluation criteria. On AV-Deepfake1M, leading models like UMMAFormer and BA-TFD+ suffer performance drops due to greater data diversity and shorter segments. In contrast, our method consistently delivers state-of-the-art results, outperforming all existing approaches across all metrics with significant margins. Specifically, DeformTrace surpasses the second-best model, DiMoDif, by 5.1%, 2.7%, 2.5%, and 4.15% in mAP@ $\{0.5, 0.75, 0.9, 0.95\}$, respectively—yielding an average improvement of 3.6%. In terms of mAR under 50, 30, 20, 10, and 5 proposals, Deform-

Type	Method	Venue	Modality	mAP(%)					mAR(%)					
				0.5	0.75	0.9	0.95	Avg.	50	30	20	10	5	Avg.
Anchor-free	Meso4	WIFS'18	\mathcal{V}	9.86	6.05	2.22	0.59	4.68	38.9	38.9	38.8	36.5	26.9	36.0
	MesoInception4	WIFS'18	\mathcal{V}	8.50	5.16	1.89	0.50	4.01	39.3	39.2	39.0	35.8	24.6	35.6
	EfficientViT	ICIAP'22	\mathcal{V}	14.7	2.42	0.13	0.01	4.32	27.1	27.0	26.4	23.9	20.3	24.9
	TriDet+VideoMAEv2	CVPR'23	\mathcal{V}	21.7	5.83	0.54	0.06	7.03	20.3	20.2	20.1	19.5	18.2	19.7
	TriDet+InternVideo	CVPR'23	\mathcal{V}	29.7	9.02	0.79	0.09	9.89	24.1	24.1	24.0	23.5	22.6	23.6
	ActionFormer+VideoMAEv2	ECCV'22	\mathcal{V}	20.2	5.73	0.57	0.07	6.65	20.0	19.9	19.8	19.1	17.8	19.3
	ActionFormer+InternVideo	ECCV'22	\mathcal{V}	36.1	12.0	1.23	0.16	12.4	27.1	27.1	27.0	26.6	25.8	26.7
	BA-TFD	DICTA'22	\mathcal{AV}	37.4	06.3	0.19	0.02	11.0	45.6	40.4	36.0	30.7	26.8	35.9
	BA-TFD+	CVIU'23	\mathcal{AV}	44.4	13.6	0.48	0.03	14.6	48.9	44.5	40.4	34.7	29.9	39.7
	UMMAFormer	MM'23	\mathcal{AV}	51.6	28.1	7.65	1.58	22.2	44.1	43.9	43.5	42.1	40.3	42.8
	DiMoDif	Arxiv'24	\mathcal{AV}	<u>86.9</u>	<u>76.0</u>	<u>28.7</u>	<u>5.43</u>	<u>49.3</u>	<u>81.6</u>	<u>80.9</u>	<u>80.3</u>	<u>78.8</u>	<u>76.6</u>	<u>79.6</u>
Query-based	TadTR	TIP'22	\mathcal{V}	60.5	33.8	5.12	1.20	25.2	57.6	55.8	54.2	51.2	50.0	53.8
	TE-TAD	CVPR'24	\mathcal{V}	71.4	54.4	11.3	2.31	34.9	66.8	64.7	64.0	61.4	59.9	63.3
	FullFormer (Ours)	–	\mathcal{AV}	77.3	57.9	22.2	4.12	40.4	70.8	67.5	66.8	65.8	63.0	66.8
	DeformTrace (Ours)	–	\mathcal{AV}	92.0	78.7	31.2	9.58	52.9	86.2	84.2	81.4	79.0	78.0	81.8

Table 2: Temporal forgery localization results on AV-Deepfake1M (Cai et al. 2024) benchmark. The modality denotes the input type: \mathcal{V} for visual, \mathcal{A} for audio and \mathcal{AV} for both. The best results are **bolded**, and second-best underlined.

Method	Total[M]	Train[M]	FLOPs[G]	Time[ms]
BA-TFD	5.5	5.5	948.1	605
BA-TFD+	152.9	152.9	<u>218.2</u>	681
UMMAFormer	165.9	49.72	1563.9	857
FullFormer (Ours)	<u>116.7</u>	<u>19.2</u>	233.7	<u>126</u>
DeformTrace (Ours)	118.3	20.8	212.4	104

Table 3: Model size and computation comparison of \mathcal{AV} TFL methods. *Total*, *Train*, and *Time* denote total parameters, trainable parameters, and per-video inference time.

Trace achieves gains of 4.6%, 3.3%, 1.1%, and 1.4%, averaging a 2.2% improvement over DiMoDif. These results clearly demonstrate the model’s strength in precisely localizing fine-grained forgery segments, even under challenging conditions.

Efficiency results. We compare the model size and computational cost of audio-visual TFL methods, reporting both total and trainable parameters to consider differences in feature extractors. Computation cost and inference time includes feature extraction. As shown in Table 3, our model is more efficient and lightweight: it reduces trainable parameters by 28.92M and 132.1M compared to UMMAFormer and BA-TFD+, and achieves the lowest GFLOPs—6.4× less than UMMAFormer. It also runs 7.3× faster than UMMAFormer and 5.8× faster than BA-TFD and BA-TFD+. Compared to our Baseline, DeformTrace slightly increases parameter count but significantly reduces computation and inference time. This efficiency benefits from the query-based architecture (fully end-to-end without post-processing) and the linear complexity of SSMs. The model’s computational cost is mainly concentrated in the sequence processing modules, where the linear-complexity SSMs account for approximately a 4:1 cost ratio between DS-SSM and DC-SSM, while the lightweight MLP-based offset prediction has a negligible contribution. Overall, DeformTrace offers a better trade-off between performance and efficiency.

#	DS-SSM	DC-SSM	\mathcal{L}_{enh}	$\mathcal{L}_{\text{coop}}$	mAP	mAR	AUC
<i>Vanilla SSM</i>							
1	\times	\times	\times	\times	41.2	68.7	90.3
<i>Deformable SSM</i>							
2	\checkmark	\times	\times	\times	44.4	73.2	95.7
3	\times	\checkmark	\times	\times	47.8	75.7	97.3
4	\checkmark	\checkmark	\times	\times	49.8	78.2	98.1
<i>Relay Token Mechanism</i>							
5	\checkmark	\checkmark	\checkmark	\times	<u>51.4</u>	<u>79.2</u>	98.3
6	\checkmark	\checkmark	\times	\checkmark	51.7	78.9	98.5
7	\checkmark	\checkmark	\checkmark	\checkmark	52.9	81.8	99.2

Table 4: Ablation study results on AV-Deepfake1M.

Ablation Study

Ablation studies on DeformTrace. To assess each component’s impact, we conduct ablations on the AV-Deepfake1M full test set, as shown in Table 4. Compared to the baseline (Line 1) that replaces all Deformable SSMs with vanilla SSMs and removes relay losses, DeformTrace (Line 7) improves mAP by 28.4%, mAR by 19.1%, and AUC by 9.9%, highlighting the effectiveness of all modules. Specifically, Lines 2 and 3 show that both DS-SSM and DC-SSM enhance performance, with DC-SSM providing larger gains due to: 1) without relay tokens, DS-SSM suffers more from long-range decay, and 2) DC-SSM enhances query–feature sequence interaction over vanilla SSM, which is crucial for proposal refinement. Line 4 shows that combining both modules boosts performance further. Comparing Line 4 with Lines 5–6 indicates that each relay loss brings moderate gains, while using both (Line 7) yields the best results, surpassing Line 4 by 3.1% mAP, 3.6% mAR, and 1.1% AUC.

Ablation studies on hyperparameters. We conduct an ablation study on the number of relay tokens N_r , using the AV-Deepfake1M full test set with an average video duration of 9 seconds. As shown in Figure 4(a), when $N_r = 1$ or 2, performance is comparable to or worse than the baseline

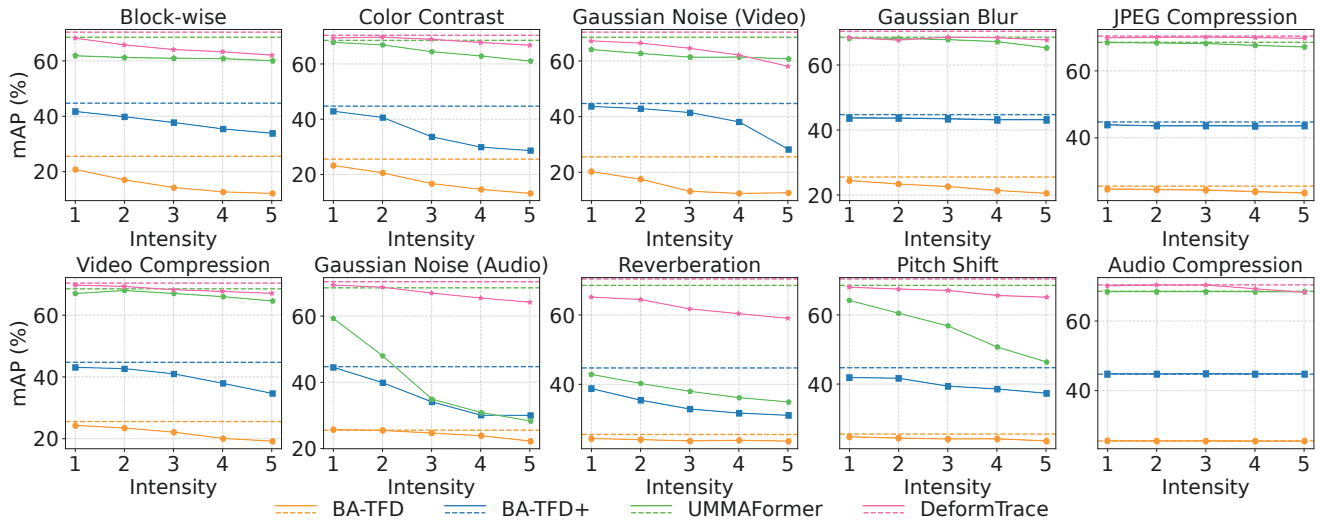


Figure 3: Robustness evaluation under various compression and degradation scenarios. The experiments include 6 visual distortions (Block-wise, Color Contrast, Gaussian Noise (video), Gaussian Blur, JPEG Compression and Video Compression) and 4 audio distortions (Gaussian Noise (audio), Reverberation, Pitch Shift and Audio Compression). In the figure, colors denote methods; solid lines show mAP across intensities, dashed lines show mAP on clean videos. DeformTrace achieves the highest mAP on clean videos and demonstrates strong robustness across various distortion scenarios.

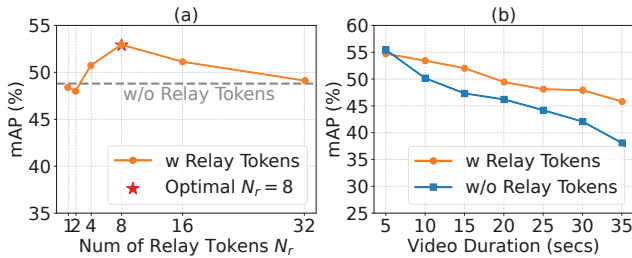


Figure 4: (a) Ablation study on the number of relay tokens (avg. video length: 9s). (b) Performance vs. video duration.

without relay tokens, as each sub-state space still contains too many tokens to alleviate long-range decay. As N_r increases, mAP improves significantly, peaking at $N_r = 8$, where token distribution per subspace aligns well with decoder queries N_q . Beyond this, further increasing N_r leads to performance drop due to over-segmentation, which impedes global information flow and raises complexity.

Performance vs. Video Duration. We select seven specific durations ranging from 5 to 35 seconds and construct sub-datasets by sampling 1,000 videos closest to each target duration from AV-Deepfake1M. This setup allows us to systematically examine how model performance varies with sequence length. As shown in Figure 4(b), model with relay tokens outperform its counterparts as duration increases, demonstrating a stronger capacity to capture long-range dependencies. In contrast, model without relay tokens shows marginally better performance on short videos, but their performance deteriorates rapidly on longer sequences. The results highlight the importance of relay tokens in maintaining

robust temporal reasoning over time.

Robustness Study

To evaluate the robustness of our model, we sampled 2,500 videos (2000 from LAV-DF, 500 from AV-Deepfake1M) and applied 10 common transmission perturbations (6 visual, 4 audio) at 5 intensity levels each, yielding 125,000 degraded videos. We tested our model alongside three audio-visual TFL methods, comparing performance using mAP. As shown in Figure 3, DeformTrace achieves the highest mAP on clean videos and exhibits strong robustness against Color Contrast, Gaussian Blur, Pitch Shift and three compression types. While performance decreases at high intensities for other distortions, our approach consistently outperforms baselines, demonstrating superior robustness.

Conclusion and Discussion

We propose DeformTrace, a novel TFL framework that integrates State Space Models (SSMs) for a better performance-efficiency trade-off. It includes a Deformable Self-SSM (DS-SSM) with dynamic receptive fields for precise boundary localization, a Deformable Cross-SSM (DC-SSM) for modeling cross-sequence interactions in sparse forgeries, and a Relay Token Mechanism to mitigate long-range information decay. By combining the global modeling of Transformers with the efficiency of SSMs, DeformTrace achieves precise localization with significantly faster inference, exceeding previous methods in both precision and robustness.

While DC-SSM is used in our work solely for modeling between forgery queries and features, it can generalize to any two independent sequences such as audio and video for audio-visual correspondence learning, enabling broader applications beyond temporal forgery localization.

Acknowledgments

This research is funded in part by the National Natural Science Foundation of China (62171326, 62371350, 62471343), Key Science and Technology Research Project of Xinjiang Production and Construction Corps (2025AB029) in 2025, Guangdong OPPO Mobile Telecommunications Corp. and Wuhan University Supercomputing Center.

References

- Afouras, T.; Chung, J. S.; and Zisserman, A. 2018. LRS3-TED: A Large-Scale Dataset for Visual Speech Recognition. arXiv:1809.00496.
- Ali, A.; Zimmerman, I.; and Wolf, L. 2025. The Hidden Attention of Mamba Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 1516–1534.
- Bai, Y.; Wang, Y.; Tong, Y.; Yang, Y.; Liu, Q.; and Liu, J. 2020. Boundary Content Graph Neural Network for Temporal Action Proposal Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, 121–137. Cham: Springer Nature Switzerland.
- Cai, Z.; Ghosh, S.; Adatia, A. P.; Hayat, M.; Dhall, A.; Gedeon, T.; and Stefanov, K. 2024. AV-Deepfake1M: A Large-Scale LLM-Driven Audio-Visual Deepfake Dataset. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, 7414–7423.
- Cai, Z.; Ghosh, S.; Dhall, A.; Gedeon, T.; Stefanov, K.; and Hayat, M. 2023. Glitch in the Matrix: A Large Scale Benchmark for Content Driven Audio-Visual Forgery Detection and Localization. *Computer Vision and Image Understanding*, 236: 103818.
- Cai, Z.; Stefanov, K.; Dhall, A.; and Hayat, M. 2022. Do You Really Mean That? Content Driven Audio-Visual Deepfake Dataset and Multimodal Method for Temporal Forgery Localization. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 1–10.
- Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *Proceedings of the European conference on computer vision (ECCV)*, 213–229. Springer.
- Chen, G.; Huang, Y.; Xu, J.; Pei, B.; Chen, Z.; Li, Z.; Wang, J.; Li, K.; Lu, T.; and Wang, L. 2024. Video Mamba Suite: State Space Model as a Versatile Alternative for Video Understanding. arXiv:2403.09626.
- Cheng, F.; and Bertasius, G. 2022. Tallformer: Temporal Action Localization with a Long-Memory Transformer. In *Proceedings of the European conference on computer vision (ECCV)*, Lecture Notes in Computer Science, 503–521. Cham: Springer Nature Switzerland.
- Cheng, H.; Liu, H.; Cai, J.; and Ye, L. 2025. CLFormer: a cross-lingual transformer framework for temporal forgery localization. *Visual Intelligence*, 3(1): 13.
- Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. VoxCeleb2: Deep Speaker Recognition. In *Proceedings of Interspeech 2018*, 1086–1090. ISCA.
- Dao, T.; and Gu, A. 2024. Transformers are SSMS: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.
- Feng, C.; Chen, Z.; and Owens, A. 2023. Self-Supervised Video Forensics by Audio-Visual Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10491–10503.
- Gu, A.; and Dao, T. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the 1st International Conference on Language Modeling (ICLM)*.
- Guo, H.; Guo, Y.; Zha, Y.; Zhang, Y.; Li, W.; Dai, T.; Xia, S.-T.; and Li, Y. 2025. MambaIRv2: Attentive State Space Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 28124–28133.
- Haliassos, A.; Ma, P.; Mira, R.; Petridis, S.; and Pantic, M. 2023. Jointly Learning Visual and Auditory Speech Representations from Raw Data. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.
- Hatamizadeh, A.; and Kautz, J. 2025. MambaVision: A Hybrid Mamba-Transformer Vision Backbone. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 25261–25270.
- Hu, J.; Zheng, J.; Wei, J.; Zhang, J.; and Stiefelhagen, R. 2024. Deformable Mamba for Wide Field of View Segmentation. arXiv:2411.16481.
- Huang, J.; Yuan, X.; Lam, C.-T.; Im, S.-K.; Lei, F.; and Bi, X. 2025. TransHFC: Joints Hypergraph Filtering Convolution and Transformer Framework for Temporal Forgery Localization. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Kim, H.-J.; Hong, J.-H.; Kong, H.; and Lee, S.-W. 2024. TE-TAD: Towards Full End-to-End Temporal Action Detection via Time-Aligned Coordinate Expression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18837–18846.
- Koutlis, C.; and Papadopoulos, S. 2024. DiMoDif: Discourse Modality-information Differentiation for Audio-visual Deepfake Detection and Localization. arXiv:2411.10193.
- Lin, C.; Xu, C.; Luo, D.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; and Fu, Y. 2021. Learning Salient Boundary Feature for Anchor-Free Temporal Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3320–3329.
- Lin, T.; Liu, X.; Li, X.; Ding, E.; and Wen, S. 2019. BMN: Boundary-Matching Network for Temporal Action Proposal Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3889–3898.
- Lin, T.; Zhao, X.; Su, H.; Wang, C.; and Yang, M. 2018. BSN: Boundary Sensitive Network for Temporal Action Proposal Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, 3–21. Cham: Springer Nature Switzerland.

- Liu, L.; Zhang, M.; Yin, J.; Liu, T.; Ji, W.; Piao, Y.; and Lu, H. 2025. DefMamba: Deformable Visual State Space Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8838–8847.
- Liu, S.; Sui, L.; Zhang, C.-L.; Mu, F.; Zhao, C.; and Ghanem, B. 2024a. Harnessing Temporal Causality for Advanced Temporal Action Detection. arXiv:2407.17792.
- Liu, X.; Wang, Q.; Hu, Y.; Tang, X.; Zhang, S.; Bai, S.; and Bai, X. 2022. End-to-end Temporal Action Detection with Transformer. *IEEE Transactions on Image Processing*, 31: 5427–5441.
- Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024b. VMamba: Visual State Space Model. *Advances in Neural Information Processing Systems*, 37: 103031–103063.
- Long, F.; Yao, T.; Qiu, Z.; Tian, X.; Luo, J.; and Mei, T. 2019. Gaussian Temporal Awareness Networks for Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 344–353.
- Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.
- Oorloff, T.; Koppiseti, S.; Bonettini, N.; Solanki, D.; Colman, B.; Yacoob, Y.; Shahriyari, A.; and Bharaj, G. 2024. AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27102–27112.
- Patro, B. N.; and Agneeswaran, V. S. 2024. SiMBA: Simplified Mamba-Based Architecture for Vision and Multivariate Time series. arXiv:2403.15360.
- Pei, X.; Huang, T.; and Xu, C. 2025. EfficientVMamba: Atrous Selective Scan for Light Weight Visual Mamba. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 6443–6451.
- Pérez-Vieites, D.; Moreira-Pérez, J. J.; Aragón-Kifute, Á.; Román-Sarmiento, R.; and Castro-González, R. 2024. Vigo: Audiovisual Fake Detection and Segment Localization. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, 11360–11364.
- Smeu, S.; Boldisor, D.-A.; Oneata, D.; and Oneata, E. 2025. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18815–18825.
- Tan, J.; Tang, J.; Wang, L.; and Wu, G. 2021. Relaxed Transformer Decoders for Direct Action Proposal Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 13526–13535.
- Wang, J.; Yan, J. N.; Gu, A.; and Rush, A. M. 2023. Pre-training Without Attention. In *Findings of the Association for Computational Linguistics: EMNLP*, 58–69. Singapore: Association for Computational Linguistics.
- Wang, P.; Lu, S.; Dai, C.; Dai, S.; and Guo, B. 2025. Temporal Action Localization with State-Sensitive Mamba and Centroid Sequences Enhancement. *Neurocomputing*, 620: 129246.
- Yang, W.; Zhou, X.; Chen, Z.; Guo, B.; Ba, Z.; Xia, Z.; Cao, X.; and Ren, K. 2023. AVoid-DF: Audio-Visual Joint Learning for Detecting Deepfake. *IEEE Transactions on Information Forensics and Security*, 18: 2015–2029.
- Ye, Z.; Xia, K.; Fu, Y.; Dong, X.; Hong, J.; Yuan, X.; Diao, S.; Kautz, J.; Molchanov, P.; and Lin, Y. C. 2025. Long-Mamba: Enhancing Mamba’s Long-Context Capabilities via Training-Free Receptive Field Enlargement. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.
- Yu, Y.; Liu, X.; Ni, R.; Yang, S.; Zhao, Y.; and Kot, A. C. 2023. PVASS-MDD: Predictive Visual-Audio Alignment Self-Supervision for Multimodal Deepfake Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34: 6926–6936.
- Zhang, C.; Wu, J.; and Li, Y. 2022. Actionformer: Localizing Moments of Actions with Transformers. In *Proceedings of the European conference on computer vision (ECCV)*, Lecture Notes in Computer Science, 492–510. Cham: Springer Nature Switzerland.
- Zhang, R.; Wang, H.; Du, M.; Liu, H.; Zhou, Y.; and Zeng, Q. 2023. UMMAFormer: A Universal Multimodal-adaptive Transformer Framework For Temporal Forgery Localization. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 8749–8759.
- Zhang, Y.; Miao, C.; Luo, M.; Li, J.; Deng, W.; Yao, W.; Li, Z.; Hu, B.; Feng, W.; Gong, T.; et al. 2024. Mfms: Learning modality-fused and modality-specific features for deepfake detection and localization tasks. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, 11365–11369.
- Zhou, Y.; and Lim, S.-N. 2021. Joint Audio-Visual Deepfake Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 14780–14789.
- Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.
- Zhu, X.; Wang, S.; Yang, J.; Yang, Y.; Tu, W.; and Wang, Z. 2025. Query-Based Audio-Visual Temporal Forgery Localization with Register-Enhanced Representation Learning. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, 8547–8556.