MG-RWKV: Multi-Grained Context-Aware RWKV for Temporal Forgery Localization

Jingchen Ni[1*], Cangjin Yu[2*], Dan Jiang[1], Quan Zhang[1], Keyu Lv[1], Shannan Yan[1], Linyue Pan[1], Ke Zhang[2†], and Chun Yuan[1†]

[1]Tsinghua University, China [2]Soochow University, China
njc24@mails.tsinghua.edu.cn, zhangke@suda.edu.cn,
yuanc@sz.tsinghua.edu.cn

Abstract. Driven by Artificial Intelligence-Generated Content (AIGC), the authenticity of audio-visual content is facing severe challenges. Temporal Forgery Localization (TFL) aims to precisely identify manipulated segments within untrimmed sequences. However, existing methods are limited by CNNs' local receptive fields or Transformers' quadratic complexity, while emerging linear models often struggle to balance global authentic context compression with local abrupt forgery perception. To address this, we propose MG-RWKV, a multi-granularity framework that leverages the data-dependent state evolution of RWKV to achieve efficient full-sequence processing with $\mathcal{O}(T)$ complexity. Our framework features three core innovations: (1) a **Bidirectional RWKV** architecture that captures bidirectional temporal contexts without quadratic overhead; (2) a **Multi-Granularity Mixture of Experts (MG-MoE)** that performs dynamic routing over explicit temporal receptive fields, adaptively selecting granularities based on forgery duration to significantly enhance decision interpretability; and (3) **Cross-Granularity Consistency (CGC)**, which aligns adjacent feature pyramid levels through hierarchical scale-wise pairing and spatial boundary-aware weighting, effectively reducing false positives in authentic regions. Extensive experiments on Lav-DF, TVIL, and Psynd datasets demonstrate that MG-RWKV achieves state-of-the-art performance with low computational cost.

Keywords: Temporal Forgery Localization · RWKV · Mixture of Experts

1 Introduction

Digital content forgery detection has long stood as a pivotal focus in multimedia security [12, 35]. Traditional forgery techniques primarily involve manipulating image data. With the rapid rise of Artificial Intelligence-Generated Content (AIGC) [24, 34, 43], however, deepfake-driven audio-visual forgeries have

* Equal contribution. † Corresponding authors.

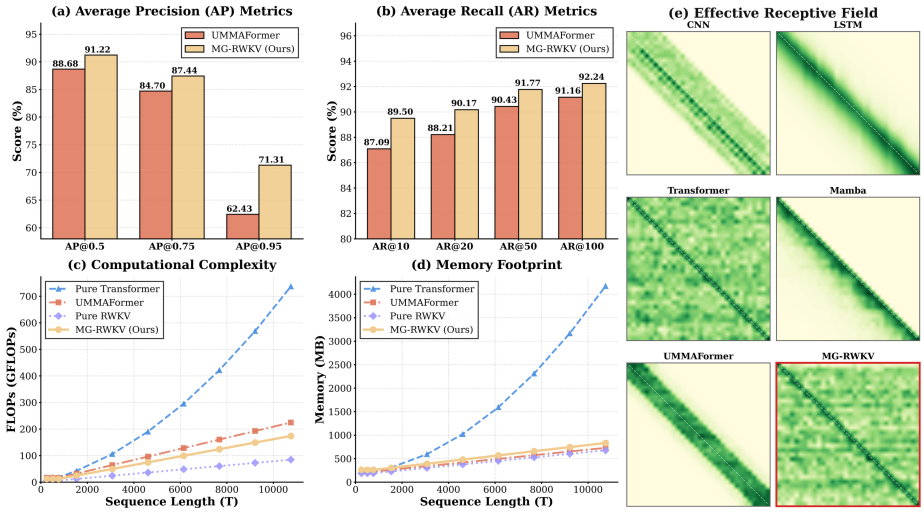


Fig. 1: Performance and computational efficiency comparison on TVIL dataset. (a) Average Precision at different thresholds. (b) Average Recall at different proposal numbers. (c) Computational complexity (FLOPs) versus sequence length. (d) Memory footprint versus sequence length. (e) Effective Receptive Field (ERF) comparison across architectures—MG-RWKV exhibits dense, long-range temporal connectivity comparable to full Transformers while maintaining linear complexity. MG-RWKV achieves superior performance with linear $\mathcal{O}(T)$ scaling.

emerged as the mainstream. The proliferation of such high-fidelity deceptive content raises severe societal concerns, underscoring the urgency for advanced detection technologies. Early detection approaches centered predominantly on facial forgeries [18, 31, 42]. In complex audio-visual scenarios, attackers often manipulate specific content segments through voice cloning or video tampering, producing highly deceptive material that poses significant challenges to traditional binary classification paradigms.

To address this gap, recent research redefines the task as Temporal Forgery Localization (TFL) [5, 17], aiming to spatially and temporally localize forged segments within untrimmed sequences. This requires the model to identify subtle manipulation traces—such as semantic replacement, emotional inconsistency, or object restoration errors—across hundreds or thousands of frames.

However, existing TFL methods face a fundamental architectural bottleneck when modeling long-range dependencies. CNN-based approaches suffer from limited receptive fields, struggling to capture global temporal inconsistencies across time spans. Conversely, Transformer-based frameworks, while possessing global context, incur quadratic $\mathcal{O}(T^2)$ complexity via self-attention, leading to severe computational and memory bottlenecks when processing full sequences. To mitigate this, some methods [51] adopt local window attention, which inevitably sacrifices global modeling capabilities. Recently, emerging linear-complexity archi-

tectures, such as state space models (e.g., Mamba [14]) and linear attention [25], have shown promise. Yet, the TFL task poses a unique requirement: the model must efficiently compress the global authentic context while remaining highly sensitive to abrupt, instantaneous boundary changes caused by forgeries. Conventional linear models often struggle to achieve this optimal balance between "global smooth compression" and "local abrupt perception". We observe that the "data-dependent decay" and dynamic state evolution mechanisms inherent in the RWKV [32] architecture naturally align with this requirement, offering an ideal paradigm for TFL.

In this paper, we propose MG-RWKV, a linear-complexity framework systematically tailored for temporal forgery localization. MG-RWKV maintains linear $\mathcal{O}(T)$ scaling while enhancing the transparency and interpretability of the localization process through three synergistic modules. First, accurate forgery boundary localization depends on both "before" and "after" contexts. Building upon traditional RWKV, we design a **Bidirectional RWKV (BiDir)** architecture that simultaneously captures past and future temporal dependencies, achieving a true global receptive field without the computational burden of Transformers. As shown in Fig. 1(e), the Effective Receptive Field analysis confirms that MG-RWKV achieves dense, long-range temporal connectivity comparable to full Transformers while maintaining linear $\mathcal{O}(T)$ complexity.

Second, forgery patterns exhibit significant variations in temporal scale—ranging from instantaneous frame flickers requiring fine-grained perception to large-scale scene generations demanding a macroscopic coarse-grained view. We design a **Multi-Granularity Mixture of Experts (MG-MoE)** module. Unlike standard black-box routing, our "experts" are constructed from convolutional branches with different dilation rates, representing temporal receptive fields with explicit physical meanings. Through input-aware dynamic routing, the model adaptively selects the appropriate granularity based on the specific forgery duration, substantially enhancing the interpretability of the decision-making process.

Finally, to address the issue of multi-scale features producing inconsistent predictions in authentic regions—a primary source of false positives—we propose the **Cross-Granularity Consistency (CGC)** constraint. CGC achieves precise feature alignment through two core designs: structurally, it performs *hierarchical scale-wise pairing* between adjacent FPN levels; and spatially, it applies *boundary-aware weighting* to relax constraints at transition frames where scale-dependent differences carry genuine semantic meaning. This design effectively aligns cross-granularity representations and sharpens temporal boundary localization accuracy.

In summary, our contributions are as follows:

- We propose the novel MG-RWKV framework, systematically exploring the application of a data-dependent linear recurrent architecture for the TFL task. This effectively breaks the efficiency-accuracy trade-off bottleneck of existing methods, distinguishing our approach from both Transformers and generic linear models.

- We design a Bidirectional RWKV architecture to capture bidirectional context and innovatively propose the MG-MoE module, which leverages dynamic routing over explicit temporal receptive fields to achieve adaptive and highly interpretable multi-scale perception.
- We introduce the CGC module, which significantly reduces false positives and improves boundary precision by cleanly integrating hierarchical cross-scale alignment and spatial boundary-aware weighting.

2 Related Work

2.1 Image Forgery Detection

Traditional IFD methods rely on handcrafted features such as color filter arrays, photo-response non-uniformity noise, illumination, and JPEG artifacts. Although effective in some cases, these methods struggle against advanced forgeries where manipulated regions blend seamlessly with the background. Recently, deep learning-based approaches [7, 11, 16, 19, 22, 29, 49] have achieved remarkable progress. For instance, MVSS-Net [11] adopts a dual-stream architecture to jointly model noise and boundary cues, PSCC-Net [22] performs bidirectional feature aggregation, and TruFor [16] fuses RGB and noise-sensitive fingerprints using a Transformer-based structure for robust trace extraction.

2.2 Temporal Forgery Localization

With the proliferation of tampered audio-visual content, accurately localizing the temporal span of forgery remains a major challenge due to data scarcity and high realism of synthetic content. To tackle this, researchers have developed benchmark datasets such as Lav-DF [5] and TVIL [51] and proposed representative models. BA-TFD [5] employs dual 3D CNN encoders with contrastive and boundary matching losses to capture modal desynchronization. UMMAFormer [51] introduces a Transformer-based temporal anomaly attention module and cross-attention feature pyramid network for long-range dependency modeling. More broadly, advances in cross-modal representation learning [28] and language-guided localization [38] underscore the importance of aligning heterogeneous cues for precise localization.

2.3 Temporal Action Detection

TAD aims to identify and localize actions in untrimmed videos. Existing approaches fall into two categories: two-stage and one-stage methods. Two-stage frameworks [13, 40] generate and classify action proposals separately, predicting action boundaries or using anchor-based strategies, but suffer from high complexity and limited end-to-end optimization. In contrast, one-stage methods [4, 21] jointly perform localization and classification within a unified network, achieving improved efficiency though still facing a performance gap compared with recent Transformer-based models. Beyond full supervision, weakly- and unsupervised methods [39, 47, 48, 50] further reduce annotation cost.

2.4 Efficient Sequence Models

Several architectures have been proposed to replace the quadratic self-attention of Transformers with linear-complexity alternatives. Linear attention methods such as Performer [8] and Linformer [37] approximate full attention through kernel tricks or low-rank projections, but often sacrifice modeling capacity for long-range dependencies. State Space Models (SSMs), notably S4 [15] and Mamba [14], reformulate sequence modeling as a selective state space recurrence, achieving $\mathcal{O}(T)$ complexity with competitive performance on sequence tasks. However, SSMs are designed with fixed or data-independent state transition mechanisms, which may limit their sensitivity to the subtle, locally-concentrated anomalies characteristic of forgery boundaries. RWKV [32] combines the efficiency of recurrent inference with data-dependent decay and in-context state modulation, providing stronger adaptive capacity for detecting temporal anomalies. Empirically, we observe that RWKV-7 outperforms Mamba on TFL (82.43 vs. 80.15 mAP on Lav-DF; see Sec. 4.4), which we attribute to its more expressive state modulation mechanism. Importantly, MG-RWKV is not merely a substitution of Transformers with RWKV—it exploits RWKV's recurrent structure to design a dilation-based multi-granularity architecture that is not naturally supported by attention-based models.

3 Methodology

3.1 Overview

Given a feature sequence $\mathbf{X} \in \mathbb{R}^{T \times D}$ of an untrimmed video, where T is the number of time steps and D is the feature dimension, the goal of Temporal Forgery Localization (TFL) is to detect forged temporal segments. Following the anchor-free detection paradigm [51], our model produces dense predictions: classification scores $\mathbf{P} \in \mathbb{R}^{T \times N_c}$ and boundary offsets $\mathbf{O} \in \mathbb{R}^{T \times 2}$ for each time position, which are then converted into segment proposals $\{(t_s^i, t_e^i, s^i)\}_{i=1}^{N}$ through post-processing.

As illustrated in Fig. 2, MG-RWKV first extracts visual and audio features using pre-trained TSN [36] and BYOL-A [30], which are fused and projected to form the input sequence $\mathbf{X}$. The sequence is then processed by L stacked MG-RWKV blocks to produce hierarchical multi-scale features $\{\mathbf{H}^{(l)}\}_{l=1}^{L}$, where each block applies dilated multi-scale convolution and bidirectional RWKV with MG-MoE routing. A top-down Feature Pyramid Network (FPN) further refines and fuses these features into $\{\mathbf{F}^{(l)}\}_{l=1}^{L}$, upon which classification and regression heads output dense forgery scores and boundary offsets. Finally, we apply Soft-NMS [3] to convert dense predictions into the top-100 segment proposals.

The framework incorporates three core innovations. The **Bidirectional RWKV Architecture** replaces quadratic self-attention with a linear-complexity recurrent mechanism, while the bidirectional scan provides global temporal context essential for boundary localization. The **Multi-Granularity Mixture of Experts (MG-MoE)** treats BiRWKV branches with structurally distinct dila-

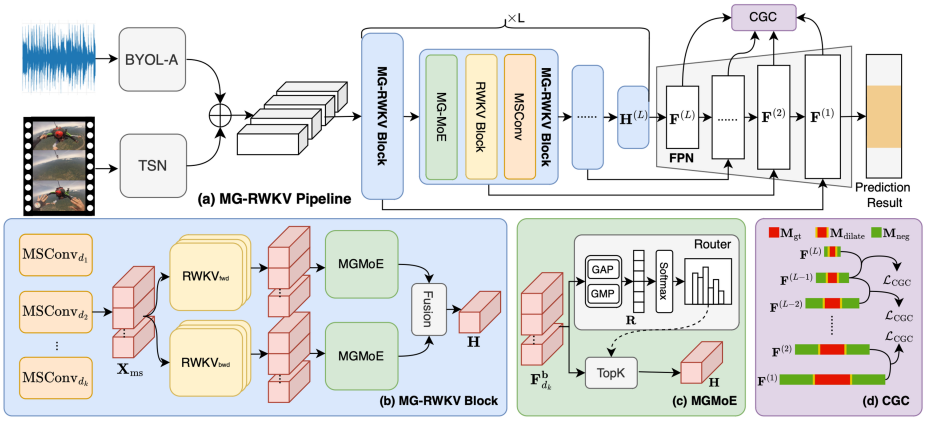


Fig. 2: Overview of the proposed MG-RWKV framework. (a) Overall pipeline with BYOL-A/TSN extractors, MG-RWKV blocks, FPN, and prediction heads. (b) MG-RWKV block with multi-scale convolutions, bidirectional RWKV, and MG-MoE. (c) MG-MoE with dynamic routing via GAP/GMP and Top-K expert selection. (d) CGC for cross-granularity alignment with boundary-aware weighting, where red denotes forged regions, yellow the dilated boundary margin, and green the negative regions.

tion rates as interpretable experts, enabling position-adaptive granularity selection through dynamic routing. The **Cross-Granularity Consistency (CGC)** enforces cross-scale feature agreement in authentic regions, resolving the cross-granularity contradictions that multi-scale modeling inherently introduces. These three components form a closed-loop design: BiDir establishes the global context upon which MG-MoE performs adaptive multi-scale routing, while CGC eliminates the inter-scale inconsistencies that such routing would otherwise introduce.

3.2 RWKV-7 for Temporal Modeling

Accurate boundary localization in TFL requires both past and future context: a unidirectional model cannot exploit post-boundary information when predicting segment starts, leading to systematically imprecise boundaries. Standard bidirectional Transformers address this via full self-attention, but incur $\mathcal{O}(T^2 \cdot d)$ complexity—for Lav-DF videos with $T \approx 1500$, this yields $\sim 2.25 \times 10^6$ pairwise computations per layer and becomes prohibitive at scale.

RWKV-7 Architecture. We adopt RWKV-7 [32], a linear-complexity recurrent model with data-dependent decay and in-context state modulation. Given input $\mathbf{x}_t \in \mathbb{R}^d$, RWKV-7 applies token shift mixing, then computes receptance $\mathbf{r}_t$, key $\mathbf{k}_t$, value $\mathbf{v}_t$, and gate $\mathbf{g}_t$ via linear projections. The key innovation generates adaptive parameters through input-dependent quadratic functions:

$$\begin{aligned} d_t &= \mathbf{w}_0 + \mathbf{w}_1 \odot \mathbf{x}'_t + \mathbf{w}_2 \odot (\mathbf{x}'_t)^2 \\ \mathbf{a}_t &= \mathbf{a}_0 + \mathbf{a}_1 \odot \mathbf{x}''_t + \mathbf{a}_2 \odot (\mathbf{x}''_t)^2 \end{aligned} \tag{1}$$

where $\mathbf{x}'_t, \mathbf{x}''_t$ are token-shifted inputs, $\mathbf{w}_i, \mathbf{a}_i$ are learnable parameters, and d_t controls the per-step decay strength. The recurrent state evolves as:

$$\mathbf{s}_t = e^{-e^{d_t}} \odot \mathbf{s}_{t-1} + \mathbf{k}_t \odot \mathbf{v}_t + \mathbf{a}_t \odot \mathbf{s}_{t-1}, \quad \mathbf{o}_t = \mathbf{r}_t \odot \mathbf{s}_t \cdot \sigma(\mathbf{g}_t) \tag{2}$$

where $e^{-e^{d_t}}$ provides numerically stable exponential decay and $\mathbf{a}_t$ enables in-context state modulation. Since $\mathbf{o}_t$ depends only on $\mathbf{x}_t$ and $\mathbf{s}_{t-1}$, the overall complexity is $\mathcal{O}(T \cdot d^2)$—linear in sequence length. Each RWKV-7 block interleaves this Time Mix module with a ReLU^2 MLP under pre-normalization and residual connections.

Bidirectional Extension. To simultaneously capture past and future context while maintaining linear complexity, we extend RWKV-7 bidirectionally by applying forward and backward scans with independent parameter sets $\boldsymbol{\theta}^{\text{fwd}}$ and $\boldsymbol{\theta}^{\text{bwd}}$. The resulting forward features $\{\mathbf{F}_k^{\text{fwd}}\}$ and backward features $\{\mathbf{F}_k^{\text{bwd}}\}$ are subsequently fused through the Multi-Granularity Mixture of Experts mechanism described in the following section.

3.3 Multi-Granularity Mixture of Experts

Video forgeries span a wide range of temporal scales: frame-level flickers demand fine-grained local analysis, while long-duration synthesis requires coarse-grained global context. Rather than treating this as an open-ended search over arbitrary resolutions, we observe that forgery temporal scales form a structured spectrum—analogous to how spatial object scales cluster around characteristic sizes in detection tasks. MG-MoE operationalizes this observation by defining each expert as a BiRWKV branch with a structurally distinct dilation rate, so the temporal receptive field of every expert is an explicit, interpretable quantity rather than an emergent property of unconstrained learned weights. The appropriate granularity varies by position, motivating a data-driven routing mechanism that selects among experts conditioned on local temporal content.

Scale-Structured Expert Bank. The forgery scale spectrum is discretized into K representative levels through dilation rates $\mathcal{D} = \{d_1, d_2, \ldots, d_K\}$. Input features $\mathbf{X} \in \mathbb{R}^{T \times C}$ are first enriched with multi-scale local context via a gated depthwise-dilated convolution:

$$\mathbf{X}_{\text{ms}} = \mathbf{X} + \gamma \cdot \text{MSConv}_{\mathcal{D}}(\mathbf{X}) \tag{3}$$

where γ is a learnable gate controlling injection strength, and $\text{MSConv}_{\mathcal{D}}$ fuses local multi-scale information across all rates in $\mathcal{D}$ before the expert split. Each branch k then processes $\mathbf{X}_{\text{ms}}$ bidirectionally at dilation d_k, yielding an effective temporal receptive field of $(w-1) \times d_k + 1$ frames, where w is the kernel size:

$$\begin{aligned} \mathbf{F}_k^{\text{fwd}} &= \text{RWKV}_{d_k}^{\text{fwd}}(\mathbf{X}_{\text{ms}}) \\ \mathbf{F}_k^{\text{bwd}} &= \text{flip}\Big(\text{RWKV}_{d_k}^{\text{bwd}}\big(\text{flip}(\mathbf{X}_{\text{ms}})\big)\Big) \end{aligned} \tag{4}$$

This yields $2K$ expert representations $\{\mathbf{F}_k^{\text{fwd}}\}_{k=1}^{K}$ and $\{\mathbf{F}_k^{\text{bwd}}\}_{k=1}^{K}$, where each expert encodes forgery evidence at a distinct temporal resolution.

Position-Adaptive Scale Selection. The routing objective is to estimate the scale preference at each temporal position from the current expert activations. Because forward and backward scans accumulate different contextual histories, they may form different scale preferences at the same position; we therefore compute independent routing weights per direction. To capture both the overall response magnitude and the presence of discriminative anomaly spikes, we represent each expert's activation by the channel-wise mean and maximum responses:

$$\begin{aligned} \mathbf{R}_{\text{mean}} &= \left[\text{Mean}_C(\mathbf{F}_k^{\text{fwd}}), \text{Mean}_C(\mathbf{F}_k^{\text{bwd}})\right]_{k=1}^{K} \in \mathbb{R}^{T\times 2K} \\ \mathbf{R}_{\text{max}} &= \left[\text{Max}_C(\mathbf{F}_k^{\text{fwd}}), \text{Max}_C(\mathbf{F}_k^{\text{bwd}})\right]_{k=1}^{K} \in \mathbb{R}^{T\times 2K} \end{aligned} \tag{5}$$

Mean pooling summarizes the broadband activation energy while max pooling preserves the most salient anomaly signals. Together they form a compact representation that captures both average response level and peak discriminative evidence—providing the router with complementary views for reliable scale selection. A lightweight 1D convolution with temperature-scaled softmax translates this into time-varying routing weights:

$$\mathbf{W}^{\mathbf{b}} = \text{softmax}\left(\frac{\text{Conv1D}([\mathbf{R}_{\text{mean}} \oplus \mathbf{R}_{\text{max}}])^{\mathbf{b}}}{\tau}\right) \in \mathbb{R}^{T\times K} \tag{6}$$

where $\mathbf{b} \in \{\text{fwd}, \text{bwd}\}$ and temperature τ governs the sharpness of scale selection. To prevent expert collapse—wherein a dense soft mixture over all scales would incentivize experts to converge toward similar average representations—we apply sparse Top-K_{top} gating, enforcing that each position activates only a subset of experts:

$$\text{TopK}(\mathbf{W}, K_{\text{top}}) = \frac{\mathbf{W} \odot \mathbb{1}_{\text{top-}K_{\text{top}}}}{\sum_k \mathbf{W}_k \odot \mathbb{1}_{\text{top-}K_{\text{top}}}} \tag{7}$$

where $\mathbb{1}_{\text{top-}K_{\text{top}}}$ retains only the K_{top} largest weights per position, enforcing specialization and maintaining representational diversity across experts. The weighted aggregation then yields direction-specific fused representations:

$$\begin{aligned} \tilde{\mathbf{W}}^{\mathbf{b}} &= \text{TopK}(\mathbf{W}^{\mathbf{b}}, K_{\text{top}}), \\ \mathbf{H}^{\mathbf{b}} &= \sum_{k=1}^{K} \tilde{\mathbf{W}}_k^{\mathbf{b}} \odot \mathbf{F}_k^{\mathbf{b}}, \quad \mathbf{b} \in \{\text{fwd}, \text{bwd}\} \end{aligned} \tag{8}$$

The two direction-specific outputs $\mathbf{H}^{\text{fwd}}$ and $\mathbf{H}^{\text{bwd}}$ are then fused via linear projection:

$$\mathbf{H} = W^{\text{fusion}}[\mathbf{H}^{\text{fwd}} \oplus \mathbf{H}^{\text{bwd}}] \tag{9}$$

where $W^{\text{fusion}} \in \mathbb{R}^{C\times 2C}$ projects the concatenated bidirectional representations back to dimension C. Setting K_{top}=2 permits adjacent granularities to be jointly activated at boundary positions, enabling smooth scale transitions that a hard top-1 selection would suppress.

3.4 Cross-Granularity Consistency

While MG-MoE captures multi-scale forgery patterns effectively, parallel branches with heterogeneous receptive fields can produce inconsistent predictions in authentic regions, elevating false positives. CGC addresses this by enforcing cosine similarity between adjacent FPN scale features exclusively in authentic regions, preserving scale-specific discriminative capacity in forged regions while suppressing cross-scale contradictions elsewhere.

Given backbone outputs $\{\mathbf{H}^{(l)}\}_{l=1}^{L}$, the FPN performs top-down fusion:

$$\begin{aligned} \mathbf{F}^{(l)} &= \mathrm{Conv}(\mathbf{H}^{(l)} + \mathrm{Upsample}(\mathbf{F}^{(l+1)})), \quad l = L-1, \ldots, 1 \\ \mathbf{F}^{(L)} &= \mathrm{Conv}(\mathbf{H}^{(L)}) \end{aligned} \tag{10}$$

The authentic region mask is constructed by dilating the ground-truth forgery mask $\mathbf{M}_{\mathrm{gt}}$ by radius r: $\mathbf{M}_{\mathrm{dilate}} = \mathrm{MaxPool1D}(\mathbf{M}_{\mathrm{gt}}, 2r{+}1)$, then taking the complement within valid positions: $\mathbf{M}_{\mathrm{neg}} = \mathbf{M}_{\mathrm{valid}} \wedge \neg\mathbf{M}_{\mathrm{dilate}}$. Boundary-aware weights $\mathbf{W}_b$ further reduce the constraint strength to 0.5 within r_b frames of segment boundaries and maintain 1.0 elsewhere, acknowledging that near-boundary frames exhibit genuine scale-dependent transition behaviors.

For adjacent FPN scale pairs $\mathcal{P} = \{(l, l{+}1)\}_{l=1}^{L-1}$, the consistency loss is:

$$\mathcal{L}_{\mathrm{CGC}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \frac{\sum_t \mathbf{M}_{\mathrm{neg}}(t) \cdot \mathbf{W}_b(t) \cdot d_{\cos}(\mathbf{F}_t^{(i)}, \mathbf{F}_t^{(j)})}{\sum_t \mathbf{M}_{\mathrm{neg}}(t)} \tag{11}$$

where $d_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$. Applying this constraint from the very first epoch risks collapsing multi-scale diversity before features have developed meaningful representations. We therefore introduce a progressive warmup schedule:

$$\lambda_{\mathrm{CGC}}(e) = \begin{cases} \lambda_0 \cdot e / E_w, & e \le E_w \\ \lambda_0, & e > E_w \end{cases} \tag{12}$$

where e is the current epoch, E_w is the warmup duration, and λ_0 is the target weight. Together, these three design dimensions of CGC reinforce each other: hierarchical scale-wise pairing propagates consistency locally between adjacent levels rather than collapsing all scales simultaneously; boundary-aware weighting relaxes constraints at transition frames where scale-dependent differences carry semantic meaning; and epoch-wise warmup defers enforcement until each scale has developed its own discriminative representation.

3.5 Training Objective

The total training loss combines classification, regression, reconstruction, and consistency objectives:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}} + \mathcal{L}_{\mathrm{reco}} + \lambda_{\mathrm{CGC}}(e)\mathcal{L}_{\mathrm{CGC}} \tag{13}$$

Table 1: Comparison with state-of-the-art methods on Lav-DF, TVIL, and Psynd datasets. AP and AR denote Average Precision and Average Recall at the specified tIoU thresholds. The best result per column is **bold**; MG-RWKV rows are highlighted in gray.

Dataset	Methods	Feature	AP@0.5	AP@0.75	AP@0.95	AR@10	AR@20	AR@50	AR@100
Lav-DF	MDS [9]	Visual	12.78	1.62	0.00	37.88	36.71	34.39	32.15
	AGT [27]	Visual	17.85	9.42	0.11	43.15	34.23	24.59	16.71
	BMN [20]	Visual	24.01	7.61	0.07	53.26	41.24	31.60	26.93
	BMN (I3D) [20]	Visual	10.56	1.66	0.00	48.49	44.39	37.13	31.55
	AVFusion [2]	Visual+Audio	65.38	23.89	0.11	62.98	59.26	54.80	52.11
	BA-TFD [5]	Visual	58.55	28.60	0.16	62.49	58.77	53.86	50.29
	BA-TFD [5]	Visual+Audio	76.90	38.50	0.25	66.90	64.08	60.77	58.42
	ActionFormer [46]	Visual	95.34	90.20	23.73	88.41	89.63	90.33	90.41
	UMMAFormer [51]	Visual	97.30	92.96	25.68	90.19	90.85	91.14	91.18
	UMMAFormer [51]	Visual+Audio	98.83	**95.54**	37.61	**92.10**	92.42	92.47	92.48
	TriDet [33]	Visual+Audio	96.29	86.84	23.64	88.69	89.71	90.39	91.00
	MFMS [52]	Visual+Audio	98.47	94.15	27.80	90.02	90.46	90.65	90.69
	ICS-AV [1]	Visual+Audio	87.40	66.80	5.72	—	—	—	—
	MG-RWKV	**Visual**	96.73	92.36	26.60	90.14	91.18	92.17	92.17
	MG-RWKV	**Visual+Audio**	**98.92**	94.81	**38.47**	91.64	**92.45**	**93.41**	**93.41**
TVIL	TAGS [26]	Visual	18.40	12.68	0.09	24.41	25.05	25.56	25.56
	DCAN [6]	Visual	82.75	75.00	3.22	64.73	66.02	68.82	69.97
	ActionFormer [46]	Visual	86.27	83.03	28.17	84.82	85.77	88.10	88.49
	UMMAFormer [51]	Visual	88.68	84.70	62.43	87.09	88.21	90.43	91.16
	MG-RWKV	**Visual**	**91.22**	**87.44**	**71.31**	**89.50**	**90.17**	**91.77**	**92.24**
Psynd	UMMAFormer [51]	Audio	100.00	**100.00**	79.87	97.60	97.60	97.60	97.60
	MG-RWKV	**Audio**	**100.00**	98.38	**90.09**	**98.61**	**98.61**	**98.61**	**98.61**

where Focal Loss $\mathcal{L}_{\text{cls}}$ handles class imbalance, DIoU Loss $\mathcal{L}_{\text{reg}}$ optimizes boundary localization, $\mathcal{L}_{\text{reco}}$ is an auxiliary reconstruction objective, and $\mathcal{L}_{\text{CGC}}$ enforces multi-scale consistency under the warmup schedule of Eq. (12).

4 Experiment

4.1 Experimental Setup

Datasets. We conduct experiments on three benchmark datasets covering diverse forgery scenarios. **Lav-DF** [5] is a multi-modal audio-visual dataset built upon VoxCeleb2 [10], featuring content-driven deepfake forgeries. **TVIL** [51] is a video-only dataset derived from YouTubeVOS 2018 [41], containing forgeries generated via video inpainting. **Psynd** [45] is an audio-only dataset based on LibriTTS [44], featuring voice cloning forgeries.

Evaluation Metrics. Following prior works [5,17], we adopt Average Precision (AP) and Average Recall (AR) as the main metrics, with the tIoU thresholds set to $\{0.5, 0.75, 0.95\}$ for AP and the Average Number of proposals (AN) set to $\{10, 20, 50, 100\}$ for AR. For Psynd, we additionally report tIoU-based results following its official protocol.

Implementation Details. Visual and audio features are extracted using pretrained TSN [36] and BYOL-A [30]. MG-RWKV adopts embedding dimension $C = 256$, pyramid blocks [2,2,5], dilation rates $\{1, 2, 4\}$, and convolution kernel

Table 2: Progressive component ablation on Lav-DF, TVIL, and Psynd datasets. Baseline (× × ×) denotes unidirectional RWKV-7 with FPN but without BiDir, MG-MoE, or CGC. ✓/× indicates whether each component is included.

Dataset	BiDir	MG-MoE	CGC	mAP	AP@0.5	AP@0.75	AP@0.95	AR@10	AR@20	AR@50	AR@100
Lav-DF	×	×	×	82.43	97.88	91.11	27.38	88.20	89.19	90.58	91.64
	✓	×	×	85.99	98.47	93.55	36.25	90.86	91.73	93.01	93.25
	✓	✓	×	86.94	98.89	94.50	37.31	91.44	92.26	93.27	93.31
	✓	✓	✓	**87.29**	**98.92**	**94.81**	**38.47**	**91.64**	**92.45**	**93.41**	**93.41**
TVIL	×	×	×	83.32	89.20	85.55	63.37	87.44	88.77	90.13	90.86
	✓	×	×	83.08	89.11	85.95	66.58	87.20	88.58	90.50	91.01
	✓	✓	×	84.35	90.43	87.42	65.87	88.38	88.99	90.65	91.57
	✓	✓	✓	**85.91**	**91.22**	**87.44**	**71.31**	**89.50**	**90.17**	**91.77**	**92.24**
Psynd	×	×	×	92.49	100.00	95.68	68.69	95.57	95.57	95.57	95.57
	✓	×	×	96.21	100.00	97.89	75.84	97.47	97.47	97.47	97.47
	✓	✓	×	97.67	100.00	98.20	87.18	98.35	98.35	98.35	98.35
	✓	✓	✓	**98.23**	**100.00**	**98.38**	**90.09**	**98.61**	**98.61**	**98.61**	**98.61**

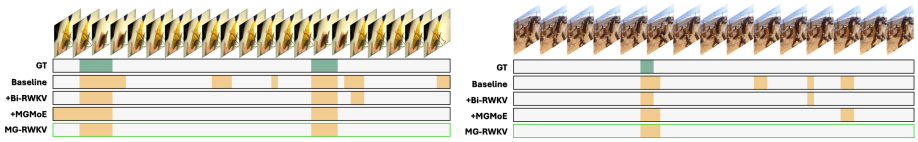


Fig. 3: Progressive component ablation on TVIL dataset. From top to bottom: Ground Truth, Baseline, +BiDir, +MG-MoE, and MG-RWKV (full). Orange indicates predicted forgery regions; green indicates authentic regions. Each component progressively improves boundary localization and reduces false positives.

size $w = 3$. MG-MoE uses temperature $\tau = 0.9$ and Top-K $K_{\text{top}} = 2$; CGC employs ignore radius $r = 8$, boundary radius $r_b = 6$, and warmup epochs $E_{\text{warmup}} = 5$. Training uses AdamW [23] with initial learning rate $\eta_0 = 10^{-4}$ and cosine annealing for 45 epochs on Lav-DF and TVIL, and 30 epochs on Psynd. Loss weights are $\lambda_{\text{reg}} = 2.0$ and $\lambda_0 = 0.01$. Data augmentation includes random cropping, label smoothing, and drop path. During inference, Soft-NMS [3] retains the top-100 proposals. All experiments are conducted on NVIDIA RTX 3090 GPUs.

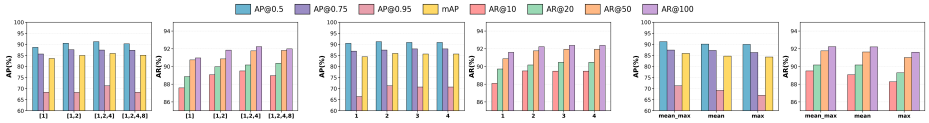
4.2 Main Experimental Results

As presented in Tab. 1, MG-RWKV achieves overall state-of-the-art performance across all three benchmark datasets, demonstrating substantial improvements over existing methods in both precision and recall metrics.

Results on Lav-DF. In visual-only mode, MG-RWKV achieves 26.60% AP@0.95 and 96.73% AP@0.5, surpassing UMMAFormer at the strictest AP@0.95 threshold while remaining comparable at looser thresholds. With audio modality, our Visual+Audio configuration reaches 38.47% AP@0.95 and 98.92% AP@0.5—both best in class—along with 93.41% AR@100, indicating that MG-RWKV maintains high recall while achieving superior boundary precision. The improvements

Table 3: Ablation study of CGC components on TVIL. $\mathcal{L}_{\text{CGC}}$: base cross-granularity consistency loss; $\mathbf{W}_b$: boundary-aware weighting (Eq. (11)); $\lambda(e)$: progressive warmup schedule (Eq. (12)).

$\mathcal{L}_{\text{CGC}}$	$\mathbf{W}_b$	$\lambda(e)$	mAP	AP@0.95	AR@100	ΔmAP
×	×	×	84.35	65.87	91.57	—
✓	×	×	84.77	68.20	91.55	+0.42
✓	✓	×	85.04	68.81	91.94	+0.27
✓	✓	✓	**85.91**	**71.31**	**92.24**	+0.87



(a) Multi-granularity scales **(b)** Top-K sparsity **(c)** Router input strategy

Fig. 4: Ablation study on MG-MoE configuration choices on the TVIL dataset. (a) Impact of different granularity scale combinations—optimal with [1,2,4]. (b) Effect of Top-K sparsity—K=2 achieves the best balance. (c) Comparison of router input strategies—mean+max pooling outperforms each alone.

are primarily driven by bidirectional context modeling, which captures both past and future temporal dependencies for more precise boundary localization.

Results on TVIL. MG-RWKV achieves 71.31% AP@0.95, 87.44% AP@0.75, and 91.22% AP@0.5, outperforming UMMAFormer by 8.88%, 2.74%, and 2.54% respectively—demonstrating consistent gains across all precision thresholds, not only at the strict boundary. The improvement stems from two synergistic mechanisms: MG-MoE dynamically selects granularity scales suited to each forgery pattern, while CGC enforces cross-scale consistency to sharpen boundary localization. The method also achieves 92.24% AR@100, maintaining strong recall alongside precision.

Results on Psynd. MG-RWKV achieves 90.09% AP@0.95, outperforming UMMAFormer by 10.22%, with near-perfect recall at 98.61% AR@100. The strong gains on audio-only forgeries—a modality that shares no visual features with the other two datasets—confirm that our multi-granularity temporal modeling generalizes well beyond visual forgery. The consistent gains across three diverse datasets spanning multi-modal deepfakes, video inpainting, and audio cloning demonstrate that MG-RWKV addresses a fundamental challenge in temporal forgery detection rather than being tuned to a specific forgery type.

4.3 Ablation Studies

Progressive Component Ablation. As shown in Tab. 2, we progressively incorporate BiDir, MG-MoE, and CGC into the baseline across three modalities. **BiDir** yields consistent AP@0.95 gains of 8.87%, 3.21%, and 7.15% on Lav-DF,

Table 4: Inference time, memory, and parameter cost of each module on Lav-DF. CGC incurs zero inference overhead as it only affects training.

Module	Time (ms)	Mem (MB)	Params (M)	mAP
Baseline	34.3±2.2	199	36.7	82.43
+BiDir	43.6±1.6	212	39.9	85.99
+MG-MoE	73.5±1.5	274	56.2	86.94
+CGC (Full)	73.4±3.4	274	56.2	**87.29**

Table 5: Comparison of linear-complexity backbones on Lav-DF.

Backbone	Time (ms)	Mem (MB)	mAP
Mamba	32.8	187	80.15
RWKV-7 (Ours)	34.3	199	**82.43**

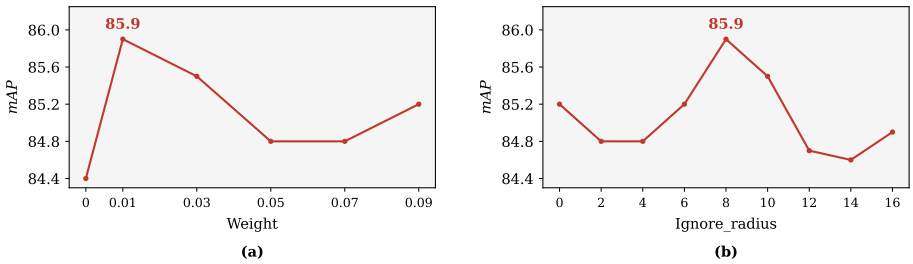


Fig. 5: CGC hyperparameter sensitivity. (a) Consistency weight λ peaks at 0.01. (b) Ignore radius r peaks at $r = 8$. Both parameters show moderate sensitivity and stable regions, validating design robustness.

TVIL, and Psynd, respectively, confirming the universal benefit of bidirectional temporal modeling. **MG-MoE** contributes mAP gains of 0.95%, 1.27%, and 1.46% on Lav-DF, TVIL, and Psynd via adaptive granularity selection. **CGC** yields the largest gains, improving mAP by 1.56% and AP@0.95 by 5.44% on TVIL, and AP@0.95 by 2.91% on Psynd, confirming that cross-scale consistency resolves boundary ambiguity. Figure 3 provides qualitative visualization of the progressive improvements.

MG-MoE Configuration Ablation. As shown in Fig. 4, **Scales** [1,2,4] achieves the best 85.91% mAP, surpassing the single-scale [1] and four-scale [1,2,4,8] at 83.65% and 85.10%, indicating that moderate granularity diversity is optimal. For **Top-K**, K=2 attains 85.91% mAP, outperforming K=1 and K=3 at 84.43% and 85.66%. For **Router Input**, combining mean and max pooling yields 85.91% mAP, exceeding the mean-only and max-only variants at 84.71% and 84.29% and confirming the complementarity of the two routing signals.

CGC Configuration Ablation. As shown in Tab. 3, the base consistency loss $\mathcal{L}_{\text{CGC}}$ yields a 0.42% mAP gain; adding boundary-aware weighting $\mathbf{W}_b$ contributes a further 0.27%; and the progressive warmup schedule $\lambda(e)$ delivers the largest gain of 0.87%, for a cumulative improvement of 1.56% mAP.

Hyperparameter Sensitivity Analysis. Figure 5 shows that consistency weight λ peaks at 0.01 with a stable range of [0.01, 0.03], and ignore radius r peaks at r=8 with a stable region of $r \in [6, 10]$. The moderate sensitivity across both parameters confirms the robustness of our CGC design.

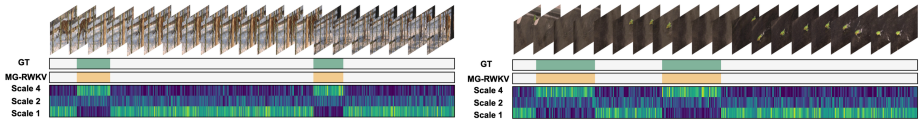


Fig. 6: MG-MoE dynamic granularity selection on TVIL. Coarse scales dominate in forged regions for broader pattern capture, while fine scales are preferred in authentic regions for precise local modeling.

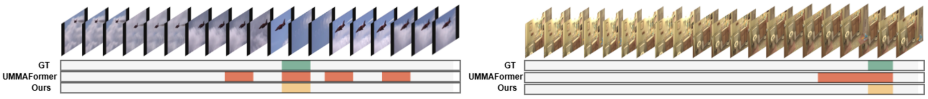


Fig. 7: Qualitative comparison on TVIL dataset. Top and bottom rows show two video samples. MG-RWKV (ours) achieves superior boundary localization and fewer false positives compared to UMMAFormer.

4.4 Efficiency and Backbone Analysis

Inference Time Ablation. As shown in Tab. 4, **BiDir** adds 9.3ms for a 3.56% mAP gain and **MG-MoE** adds 29.9ms for 0.95% mAP, while **CGC** incurs zero inference overhead. The full model achieves 87.29% mAP at 73.4ms, demonstrating a favorable efficiency-accuracy trade-off.

Linear Backbone Comparison. As shown in Tab. 5, replacing RWKV-7 with Mamba [14] under identical settings lowers mAP from 82.43% to 80.15%, confirming that RWKV's data-dependent decay and in-context state modulation are better suited for detecting locally-concentrated forgery anomalies.

4.5 Qualitative Analysis

Dynamic Granularity Selection Visualization. Figure 6 visualizes MG-MoE router weights on TVIL. Coarse-grained scales dominate in forged regions while fine-grained scales are preferred in authentic regions, with smooth transitions at boundaries confirming that the router learns position-adaptive temporal properties rather than fitting discrete labels.

Detection Result Comparison. Figure 7 compares MG-RWKV and UMMAFormer on TVIL, showing that our method achieves sharper boundary localization and fewer false positives in authentic regions.

5 Conclusion

We propose MG-RWKV, a linear-complexity framework for temporal forgery localization integrating Bidirectional RWKV, MG-MoE, and CGC. Across three benchmarks, it attains 87.29% mAP on Lav-DF and improves AP@0.95 over the previous best by 8.88% and 10.22% on TVIL and Psynd, confirming that a structured multi-scale recurrent design can match or surpass Transformer-based methods at substantially lower cost with $\mathcal{O}(T)$ complexity.

Acknowledgements

This work is supported by the SSTIC Grant (KJZD20230923115106012, KJZD20230923114916032, and GJHZ20240218113604008).

References

1. Anshul, A., Gopal, S., Rajan, D., Chng, E.S.: Intra-modal and cross-modal synchronization for audio-visual deepfake detection and temporal localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13826–13836 (2025)
2. Bagchi, A., Mahmood, J., Fernandes, D., Sarvadevabhatla, R.K.: Hear me out: Fusional approaches for audio augmented temporal action localization. arXiv preprint arXiv:2106.14118 (2021)
3. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-NMS–improving object detection with one line of code. In: Proceedings of the IEEE international conference on computer vision. pp. 5561–5569 (2017)
4. Buch, S., Escorcia, V., Ghanem, B., Fei-Fei, L., Niebles, J.C.: End-to-end, single-stream temporal action detection in untrimmed videos. In: Proceedings of the British Machine Vision Conference (BMVC). BMVA Press (2017)
5. Cai, Z., Stefanov, K., Dhall, A., Hayat, M.: Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In: 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA). pp. 1–10. IEEE (2022)
6. Chen, G., Zheng, Y.D., Wang, L., Lu, T.: DCAN: improving temporal action detection via dual context aggregation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 248–257 (2022)
7. Chen, Y., Huang, X., Zhang, Q., Li, W., Zhu, M., Yan, Q., Li, S., Chen, H., Hu, H., Yang, J., et al.: GIM: A million-scale benchmark for generative image manipulation detection and localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2311–2319 (2025)
8. Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al.: Rethinking attention with performers. In: International Conference on Learning Representations (2021)
9. Chugh, K., Gupta, P., Dhall, A., Subramanian, R.: Not made for each other-audio-visual dissonance-based deepfake detection and localization. In: Proceedings of the 28th ACM international conference on multimedia. pp. 439–447 (2020)
10. Chung, J.S., Nagrani, A., Zisserman, A.: VoxCeleb2: Deep speaker recognition. In: Interspeech. pp. 1086–1090 (2018)
11. Dong, C., Chen, X., Hu, R., Cao, J., Li, X.: MVSS-Net: Multi-view multi-scale supervised networks for image manipulation detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3539–3553 (2022)
12. El-Shafai, W., Fouda, M.A., El-Rabaie, E.S.M., El-Salam, N.A.: A comprehensive taxonomy on multimedia video forgery detection techniques: challenges and novel trends. Multimedia Tools and Applications **83**(2), 4241–4307 (2024)
13. Gao, J., Yang, Z., Nevatia, R.: Cascaded boundary regression for temporal action detection. arXiv preprint arXiv:1705.01180 (2017)
14. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: Conference on Language Modeling (COLM) (2024)

15. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. In: International Conference on Learning Representations (2022)
16. Guillaro, F., Cozzolino, D., Sud, A., Dufour, N., Verdoliva, L.: TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20606–20615 (2023)
17. He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., Liu, Z.: ForgeryNet: A versatile benchmark for comprehensive forgery analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4360–4369 (2021)
18. Huang, B., Wang, Z., Yang, J., Ai, J., Zou, Q., Wang, Q., Ye, D.: Implicit identity driven deepfake face swapping detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4490–4499 (2023)
19. Kwon, M.J., Yu, I.J., Nam, S.H., Lee, H.K.: CAT-Net: Compression artifact tracing network for detection and localization of image splicing. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 375–384 (2021)
20. Lin, T., Liu, X., Li, X., Ding, E., Wen, S.: BMN: Boundary-matching network for temporal action proposal generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3889–3898 (2019)
21. Lin, T., Zhao, X., Shou, Z.: Single shot temporal action detection. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 988–996 (2017)
22. Liu, X., Liu, Y., Chen, J., Liu, X.: PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. IEEE Transactions on Circuits and Systems for Video Technology **32**(11), 7505–7517 (2022)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
24. Lyu, S.: Deepfake the menace: mitigating the negative impacts of AI-generated content. Organizational Cybersecurity Journal: Practice, Process and People (ahead-of-print) (2024)
25. Ma, X., Zhou, C., Kong, X., He, J., Gui, L., Neubig, G., May, J., Zettlemoyer, L.: Mega: Moving average equipped gated attention. In: International Conference on Learning Representations (ICLR) (2023)
26. Nag, S., Zhu, X., Song, Y.Z., Xiang, T.: Proposal-free temporal action detection via global segmentation mask learning. In: European Conference on Computer Vision. pp. 645–662. Springer (2022)
27. Nawhal, M., Mori, G.: Activity graph transformer for temporal action localization. arXiv preprint arXiv:2101.08540 (2021)
28. Ni, J., Lyu, K., Guo, Y., Yuan, C.: Semantic alignment and hard sample retraining for visible-infrared person re-identification. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2025)
29. Ni, J., Zhang, Q., Jiang, D., Lv, K., Zhang, K., Yuan, C.: FCL-COD: Weakly supervised camouflaged object detection with frequency-aware and contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7439–7449 (2026)
30. Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., Kashino, K.: BYOL for audio: Self-supervised learning for general-purpose audio representation. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2021)
31. Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I.E., Nyameko, R., Aluvala, S., Vimal, V.: Deepfake generation and detection: Case study and challenges. IEEE Access **11**, 143296–143323 (2023)

32. Peng, B., Zhang, R., Goldstein, D., Alcaide, E., Du, X., Hou, H., Lin, J., Liu, J., Lu, J., Merrill, W., et al.: RWKV-7 "goose" with expressive dynamic state evolution. arXiv preprint arXiv:2503.14456 (2025)
33. Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J., Tao, D.: TriDet: Temporal action detection with relative boundary modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18857–18866 (2023)
34. Shoaib, M.R., Wang, Z., Ahvanooey, M.T., Zhao, J.: Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models. In: 2023 international conference on computer and applications (ICCA). pp. 1–7. IEEE (2023)
35. Tyagi, S., Yadav, D.: A detailed analysis of image and video forgery detection techniques. The Visual Computer **39**(3), 813–833 (2023)
36. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016)
37. Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H.: Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768 (2020)
38. Wang, Y., Ni, J., Liu, Y., Yuan, C., Tang, Y.: IterPrime: Zero-shot referring image segmentation with iterative Grad-CAM refinement and primary word emphasis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8159–8168 (2025)
39. Xia, R., Jiang, D., Zhang, Q., Zhang, K., Yuan, C.: CLIP-AE: CLIP-assisted cross-view audio-visual enhancement for unsupervised temporal action localization. In: 2025 IEEE International Conference on Image Processing (ICIP). pp. 2014–2018. IEEE (2025)
40. Xu, H., Das, A., Saenko, K.: R-C3D: Region convolutional 3D network for temporal activity detection. In: Proceedings of the IEEE international conference on computer vision. pp. 5783–5792 (2017)
41. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: YouTube-VOS: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)
42. Yan, Z., Zhang, Y., Fan, Y., Wu, B.: UCF: Uncovering common features for generalizable deepfake detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22412–22423 (2023)
43. Yu, X., Wang, Y., Chen, Y., Tao, Z., Xi, D., Song, S., Niu, S., Li, Z.: Fake artificial intelligence generated contents (FAIGC): A survey of theories, detection methods, and opportunities. arXiv preprint arXiv:2405.00711 (2024)
44. Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R.J., Jia, Y., Chen, Z., Wu, Y.: LibriTTS: A corpus derived from LibriSpeech for text-to-speech. arXiv preprint arXiv:1904.02882 (2019)
45. Zhang, B., Sim, T.: Localizing fake segments in speech. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 3224–3230. IEEE (2022)
46. Zhang, C.L., Wu, J., Li, Y.: ActionFormer: Localizing moments of actions with transformers. In: European Conference on Computer Vision. pp. 492–510. Springer (2022)
47. Zhang, Q., Fang, J., Qi, Y., Wan, M., Ma, G., Zhang, K., Yuan, C.: EAV-Mamba: Efficient audio-visual representation learning for weakly-supervised temporal action localization. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2025)

48. Zhang, Q., Fang, J., Yuan, R., Tang, X., Qi, Y., Zhang, K., Yuan, C.: Weakly supervised temporal action localization via dual-prior collaborative learning guided by multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24139–24148 (2025)
49. Zhang, Q., Qi, Y., Tang, X., Fang, J., Lin, X., Zhang, K., Yuan, C.: IMDPrompter: Adapting SAM to image manipulation detection by cross-view automated prompt learning. In: The Thirteenth International Conference on Learning Representations (2025)
50. Zhang, Q., Qi, Y., Tang, X., Yuan, R., Lin, X., Zhang, K., Yuan, C.: Rethinking pseudo-label guided learning for weakly supervised temporal action localization from the perspective of noise correction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10085–10093 (2025)
51. Zhang, R., Wang, H., Du, M., Liu, H., Zhou, Y., Zeng, Q.: UMMAFormer: A universal multimodal-adaptive transformer framework for temporal forgery localization. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 8749–8759 (2023)
52. Zhang, Y., Miao, C., Luo, M., Li, J., Deng, W., Yao, W., Li, Z., Hu, B., Feng, W., Gong, T., et al.: MFMS: Learning modality-fused and modality-specific features for deepfake detection and localization tasks. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 11365–11369 (2024)