This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.



MS-Temba: Multi-Scale Temporal Mamba for Understanding Long Untrimmed Videos

Arkaprava Sinha[1] **Monish Soundar Raj**[2]
Pu Wang[1] **Ahmed Helmy**[1] **Hieu Le**[1] **Srijan Das**[1]
[1] UNC Charlotte [2] UNC Chapel Hill

https://mstemba.github.io

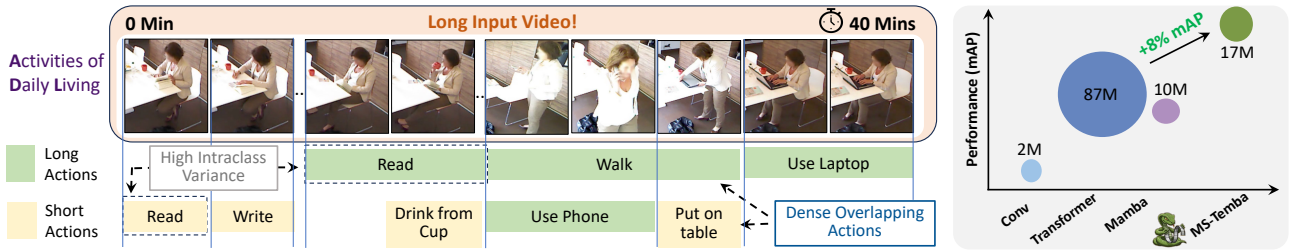


Figure 1. **Left:** Temporal Action Detection poses unique challenges, including the need for models capable of processing long video sequences, the presence of both short and long actions with high intra-class variance, and the complexity of densely overlapping actions. **Right:** Our proposed method, **MS-Temba**, achieves **state-of-the-art performance** while being **5× more parameter-efficient** compared to transformer-based approaches.

Abstract

*Temporal Action Detection (TAD) in untrimmed videos poses significant challenges, particularly for Activities of Daily Living (ADL) requiring models to (1) process long-duration videos, (2) capture temporal variations in actions, and (3) simultaneously detect dense overlapping actions. Existing CNN and Transformer-based approaches, struggle to jointly capture fine-grained detail and long-range structure at scale. State-space Model (SSM) based Mamba offers powerful long-range modeling, but naïve application to TAD collapses fine-grained temporal structure and fails to account for the challenges inherent to TAD. To this end, we propose **Multi-Scale Temporal Mamba (MS-Temba)**, which extends Mamba to TAD with newly introduced dilated SSMs. Each Temba block, comprising dilated SSMs coupled with our proposed additional losses, enables the learning of discriminative representations across temporal scales. A lightweight **Multi-scale Mamba Fuser** then unifies these multi-scale features via SSM-based aggregation, yielding precise action-boundary localization. With only 17M parameters, MS-Temba achieves state-of-the-art performance on densely labeled ADL benchmarks TSU & Charades, and further generalizes to long-form video summarization, setting new state-of-the-art results on TVSum & SumMe.*

1. Introduction

Understanding human actions in long, untrimmed videos is fundamentally harder than classifying short, curated clips [21, 34, 52]. Real-world recordings, especially in Activities of Daily Living (ADL) settings, span hours, capturing people *walking*, *eating*, *using phones*, often simultaneously and with no clear start or end boundaries. Within the same sequence, brief atomic events such as *Drink from Cup* intertwine with prolonged activities like *Use Laptop*, producing highly irregular, overlapping temporal patterns.

Temporal Action Detection (TAD) aims to make sense of this continuous stream by recognizing actions and precisely localizing when they occur. This capability is crucial to ADL applications such as patient monitoring, assistive living, and smart homes [1, 15], where systems must interpret complex, concurrent behaviors unfolding over long durations. As illustrated in Figure 1, realistic ADL videos pose three key challenges: (**Challenge 1**) capturing complex temporal dependencies in long, untrimmed videos ($\sim$ 40 min) (**Challenge 2**) representing actions spanning vastly different temporal scales, ranging from short atomic events such as *Drink from Cup* to long sustained activities like *Use Laptop*, while handling intra-class temporal variations (*e.g. Read* occurring at different temporal scales), and (**Challenge 3**) detecting densely overlapping actions, where

multiple activities (*e.g.*, *Walk* and *Use Phone*) occur simultaneously. Addressing these challenges requires models that can reason jointly over short and long range dependencies, disentangle concurrent actions and maintain precise temporal localization across complex ADL contexts.

Existing TAD frameworks, typically built on temporal CNNs or Transformers, struggle to meet these demands. Temporal CNNs capture local temporal patterns more efficiently, but their limited receptive fields hinder modeling of long-range dependencies. Transformer-based architectures [11, 62] have achieved strong performance in TAD by leveraging global self-attention, yet their quadratic complexity makes them less scalable to long video sequences. Recently, Mamba architectures based on "*state-space models*" (SSMs) [25, 26] have emerged as a promising alternative, offering linear-time information propagation across temporal sequences and enabling efficient long-range reasoning in video understanding [6, 38, 74]. However, existing SSMs [38, 46, 70] typically operate at a single temporal scale, often blending short-term motion with surrounding context and degrading fine temporal granularity. This hinders their adaptation to TAD, particularly in densely overlapping ADL scenarios.

To this end, we propose **Multi-Scale Temporal Mamba (MS-Temba)**, which extends the state-space formulation into a novel *multi-scale*, *dilated* architecture designed to capture both short- and long-term temporal dynamics within a unified framework. The core of MS-Temba lies in its **Temba Blocks**, introducing *dilated SSMs*. The key intuition is straightforward: rather than processing the video sequence with a single, fixed receptive field, we employ multiple dilated SSM branches, each operating at a distinct temporal stride. This design complements the standard SSM scanning process, which traverses all tokens sequentially (Figure 2). Short-dilation branches focus on transient, fine-grained actions, whereas long-dilation branches capture extended temporal dependencies. To ensure coherence across these asynchronous branches, we introduce a projection-based alignment that harmonizes their action predictions and mitigates scale drift. In addition, we employ scale-aware auxiliary supervision that encourages blocks to learn scale-specific, temporally discriminative representations.

Furthermore, we design a lightweight **Multi-scale Mamba Fuser (MS-Fuser)** that integrates the complementary representations learned by Temba blocks through SSM-based aggregation, producing temporally coherent and boundary-accurate predictions. This state-space formulation preserves the efficiency and modeling strength of Mamba while restoring the temporal sensitivity necessary for dense, overlapping ADL scenarios. Together, Temba and MS-Fuser enable MS-Temba to jointly model fine-grained short-term dynamics and extended long-range structures, offering a comprehensive solution for robust

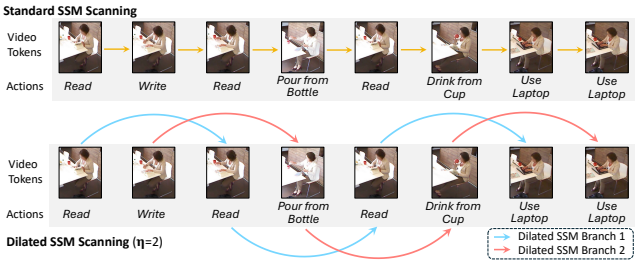


Figure 2. Illustration of Standard vs Dilated SSM scanning.

temporal action detection in complex video sequences. To assess the generalizability of MS-Temba beyond TAD, we evaluate it on video summarization, the task of condensing long videos into shorter, semantically meaningful sequences that preserve the overall narrative. MS-Temba exhibits superior performance over competitive baselines, indicating its capacity to generalize to long-form video understanding tasks.

To summarize the main contributions:

- We propose **Multi-Scale Temporal Mamba** (MS-Temba), the first Mamba-based architecture tailored to densely labeled TAD in untrimmed videos.
- We introduce **Dilated SSMs** for multi-scale temporal representation learning and design a lightweight **Multi-scale Mamba Fuser** (MS-Fuser) module that aggregates these complementary scales via SSM-based fusion.
- MS-Temba achieves state-of-the-art performance on densely labeled TAD benchmarks, TSU and Charades, with only 17M parameters, and further demonstrates strong generalization attaining state-of-the-art results on video summarization benchmarks (TVSum & SumMe).

2. Related Work

Temporal Action Detection. Understanding actions in long temporal sequences has become a key task in computer vision [13, 14, 24, 39, 68, 72]. Early action detection approaches focused on videos with sparse action labels [4, 32, 72], primarily using proposal-based methods [67, 73] inspired by object detection to generate action proposals. However, temporal action detection with densely labeled action distribution offers greater challenges due to concurrent actions, where proposal-based methods struggle with the combinatorial explosion in proposal generation. End-to-end methods [7, 60] jointly train the visual backbone and perform temporal modeling while optimizing the backbone for action detection. However, these methods are computationally prohibitive and forces the video to be broken into short windows, losing global temporal context.

This led to the adoption of temporal convolution-based models [19, 40, 41, 49] for processing long videos. Convolutional methods like PDAN [10] and TGM [50] use specialized kernels designed to model long-term dependencies and capture composite actions. However, the shared kernels in temporal convolutions are limited to neighboring frames, and even with dilations, struggle to capture global

relationships between frames. This limitation is significant for densely labeled Temporal Action Detection, where long-term actions are common. In response, transformer-based architectures [16, 18, 45, 63, 64], which have proven effective in computer vision, have been adapted for action detection [11, 22, 53, 54, 60, 62]. These architectures can model both short- and long-term dependencies in videos. For example, MLAD [62] employs a Transformer encoder to capture cross-class features and temporal relationships, while MS-TCT [11] uses a temporal hierarchy to model local relationships through temporal convolutions and global interactions through attention. These transformer based approaches enhances multi-scale representation learning, but the effectiveness comes at a high computational cost. The linear complexity of state space models [25] makes it an ideal test bed for processing long untrimmed videos.

State Space Models (SSMs). Recently, Structured State Space Model (S4) [26] has gained popularity in effectively processing long sequence lengths in linear complexity. Mamba [25] addresses some challenges of S4 (selective copying and induction heads) and introduces data-dependent SSM layer and a hardware aware efficient algorithm. Consequently, Mamba has been adapted for various computer vision downstream tasks using static images such as image classification, object detection, semantic segmentation, image restoration in [2, 27, 30, 42, 43, 66, 74]. Naturally, these state space models have been adapted for videos in a wide range of applications such as action classification [38, 46, 77], video object segmentation [69], video retrieval [61], action anticipation [70] and physiological measurement from face videos [47]. However, these methods are trained on videos which span only few seconds and have a global class per video. It is non-trivial to adapt these methods for action detection since action detectors requires two stage training, first feature extraction from untrimmed videos using a visual backbone and then temporal learning of these features for dense action-label prediction. For long-range temporal modeling, TranS4mer [31] has shown effectiveness in video classification and movie scene detection tasks by identifying scene changes in videos. MOGO [44] performs video action detection by spatially localizing and classifying atomic actions in videos. Closest to our approach, Video Mamba Suite [6] performs action localization on datasets like HACS [72] and GTEA [20]. These Mamba-based methods rely on frame sampling and compressing all information within their architectures limiting them to process videos around 3 minutes. However, in this work, we target untrimmed videos exceeding 40 minutes, featuring dense labels and co-occurring actions. State space models have not yet been applied to action detection in such untrimmed videos. To the best of our knowledge, MS-Temba is the first to leverage Mamba for the task of temporal action detection in densely labeled untrimmed videos.

3. Preliminaries

State space models map continuous input sequences $x(t) \in \mathbb{R}$ to continuous outputs $y(t) \in \mathbb{R}$ using a hidden state $h(t) \in \mathbb{R}^n$. For the above mapping and updating of the hidden state, SSMs use a transition state matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ projection matrices $\mathbf{B} \in \mathbb{R}^{n \times 1}$ and $\mathbf{C} \in \mathbb{R}^{1 \times n}$. The update equations are given by:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t) \\ y(t) &= \mathbf{C}h(t) \end{aligned} \tag{1}$$

Previous works [25, 26, 74], use discrete versions of the above system using a discretization parameter $\mathbf{\Delta}$ known as the step size. Following is an example of a discretization transformation using $\mathbf{\Delta}$, known as zero-order hold (ZOH):

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\mathbf{\Delta A}) \\ \overline{\mathbf{B}} &= (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I})\mathbf{\Delta B} \end{aligned} \tag{2}$$

This discretization of $\mathbf{A}$, $\mathbf{B}$ to $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$ leads to the following updated state-space equations:

$$\begin{aligned} h_t &= \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t \\ y_t &= \mathbf{C}h_t \end{aligned} \tag{3}$$

Standard SSMs are typically designed to process 1D inputs in a linear sequence. However, [38, 74] introduce bidirectional scanning of tokens, which are more effective for vision tasks. In this approach, the SSM operates on the input sequence v and it's reversed counterpart v', fusing the projected outputs for robust bidirectional sequence modeling. This enhancement enables more effective representation learning, especially in domains requiring complex temporal dependencies. In this work, we extend the Mamba based architectures for understanding long untrimmed videos.

4. Proposed MS-Temba

In Temporal Action Detection (TAD), we aim to classify multiple activities at each time-step of an untrimmed video. Given a video of T temporal segments, we associate an action label $\mathrm{Y} = \{y_t\}$ for each timestep t.

We introduce **Temporal Mamba (MS-Temba)**, a TAD model capable of detecting long as well as short actions efficiently in long untrimmed videos. The overall architecture of MS-Temba is illustrated in Figure 3. Our proposed MS-Temba consists of: i) a **Visual Backbone** to extract discriminative features from untrimmed videos, ii) **Temba** blocks which capture temporal information at different scales, iii) **Multi-scale Mamba Fuser** to combine the learned temporal representations across different temporal scales, and iv) a **Classification Head** to predict multiple actions within each temporal segment of the input video. In the following sections, we detail each of these components.

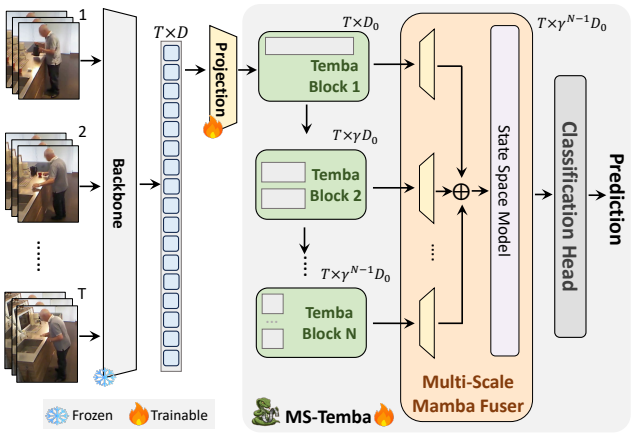


Figure 3. Overall Architecture of **Multiscale Temporal Mamba (MS-Temba)** for action detection. MS-Temba is composed of a frozen pretrained Visual Backbone, Temporal Mamba (Temba) Blocks for learning representations at multiple temporal scale spaces through dilated SSMs. The Multi-scale Mamba Fuser employs an SSM for effectively fusing the multi-scale features which is projected by the Classification Head for dense action detection.

4.1. Visual Backbone

The input to MS-Temba is an untrimmed video potentially spanning long durations (several minutes). Training an end-to-end model across both spatial and temporal dimensions poses significant computational demands. Consequently, in line with prior work [10, 11, 19, 71], we first segment the untrimmed video into T contiguous, non-overlapping segments, each consisting of 8 or 16 frames. Each segment, known as a visual token, is encoded via a visual backbone, capturing spatial features through either a 3D ConvNet (e.g., I3D [5]) pre-trained on the target dataset, or a larger foundational model such as CLIP [51]. Unlike action classification tasks, sparse frame sampling is unsuitable here, as it risks omitting short-duration actions crucial for detection. The segment-wise representations are then stacked to form an input feature vector $v \in \mathbb{R}^{B\times T\times D}$, where B is the batch of videos, D is the feature dimension and T is the total temporal sequence length. Finally, v is projected to $z_0 \in \mathbb{R}^{B\times T\times D_0}$ to obtain a latent representation of the input, which serves as input tokens to our proposed Temba Blocks.

4.2. Temporal Mamba (Temba) Blocks

TAD requires models that simultaneously capture fine-grained dynamics and long-range temporal dependencies. We introduce Temba blocks, the core computational unit of MS-Temba, which is a novel formulation of *dilated State Space Models* for multi-scale temporal representation learning. Each Temba block consists of two key components: (i) a set of *dilated SSM branches*, (ii) *scale-aware auxiliary supervision*.

Each Temba block begins with a linear projection that increases the representational capacity, enabling deeper temporal abstractions as the hierarchy progresses. The feature dimensionality increases by a factor of γ relative to the preceding block. Given the weight matrix $\mathbf{W}_k$ and bias term $\boldsymbol{\beta}_k$ for the k-th block, the transformation is defined as:

$$x_k = \mathbf{W}_k z_{k-1} + \boldsymbol{\beta}_k, \quad x_k \in \mathbb{R}^{B\times T\times \gamma D_{k-1}}, \quad k \geq 1. \tag{4}$$

This progressive expansion enables each Temba block to model increasingly complex temporal dependencies while maintaining computational efficiency. In Figure 4, we illustrate the third Temba block with $\eta = 3$.

i) *Dilated SSM*. Each Temba consists of *dilated SSM*, which allows MS-Temba to capture dependencies at varying temporal strides. Given an input feature sequence $x_k \in \mathbb{R}^{B\times T\times D_k}$, we define a non-parametric, invertible mapping $\Phi_\eta : \mathbb{R}^{B\times T\times D_k} \to \mathbb{R}^{\eta B\times \lceil T/\eta\rceil \times D_k}$, which partitions each temporal sequence into η disjoint subsequences sampled with stride η. The dilation rate η controls the temporal resolution at which each subset of tokens is processed. Concretely, letting

$$\left(X^{(0)}, X^{(1)}, \ldots, X^{(\eta-1)}\right) = \Phi_\eta(x), \tag{5}$$

each subset correspond to a dilated slice of the temporal axis:

$$X^{(i)} = \{\, x_t \mid t = i + j\eta,\ 0 \leq j < \lceil T/\eta \rceil \,\}, \quad i = 1, 2, \ldots, \eta. \tag{6}$$

In this formulation, $X^{(1)}$ contains tokens sampled at positions $1, 1+\eta, 1+2\eta, \ldots$, while $X^{(2)}$ starts from the second token and follows the same stride. As a result, Φ_η divides the sequence into η temporally distinct streams, each representing a different dilated phase of the input sequence.

Each subsequence is then processed by an individual dilated SSM branch parameterized by $(\mathbf{A}^{(i)}, \mathbf{B}^{(i)}, \mathbf{C}^{(i)})$. The output of the i-th SSM is expressed as

$$y_t^{(i)} = \mathbf{C}^{(i)} h_t^{(i)}, \quad \text{where} \tag{7}$$

$$h_t^{(i)} = \overline{\mathbf{A}^{(i)}} h_{t-1}^{(i)} + \overline{\mathbf{B}^{(i)}} X_t^{(i)} \tag{8}$$

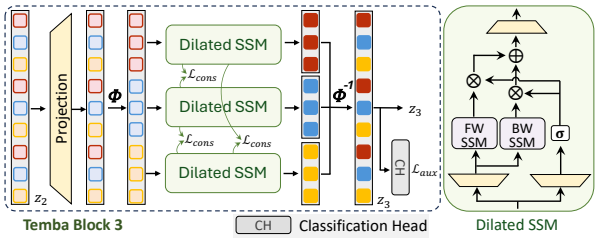


Figure 4. Temba Block 3 with $\eta = 3$. Tokens are linearly projected, grouped with stride η using Φ, processed by SSMs, aligned with $\mathcal{L}_{\text{cons}}$, and reassembled by Φ^{-1}. A classification head then produces the auxiliary loss $\mathcal{L}_{\text{aux}}$.

Distinct parameterizations across the η SSM branches allow each to learn temporal relationships at a unique dilation, resulting in diverse and complementary receptive fields.

However, to ensure coherence across neighboring temporal contexts, we introduce a **projection alignment** mechanism that enforces consistency among the output projections of dilated SSM branches within each block. Intuitively, if an action is activated in the state representation of one SSM at time t_s, adjacent temporal segments t_{s-1} and t_{s+1} should exhibit similar activation patterns in adjacent dilated SSM branches in each block. To achieve this inductive bias, we define a **pairwise consistency loss** that aligns the projection matrices across the branches. Let $\mathbf{C}_i$ and $\mathbf{C}_j$ denote the output projection matrices of the i-th and j-th dilated SSMs in a block. The pairwise consistency loss for a pair (i, j) is

$$\mathcal{L}_{\text{cons}} = 1 - \text{sim}(\widehat{\mathbf{C}}_i, \widehat{\mathbf{C}}_j), \quad i \neq j, \tag{9}$$

where $\widehat{\mathbf{C}}_i$, $\widehat{\mathbf{C}}_j$ are their fattened, ℓ_2-normalized forms, and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. This pairwise consistency loss encourages the learned state projections to remain semantically aligned across SSMs within the same dilated block, while maintaining diversity across temporal receptive fields.

After the dilated SSM branches process their respective subsequences, the outputs are reassembled to restore the original temporal ordering via the inverse mapping:

$$\Phi_\eta^{-1} : \mathbb{R}^{\eta B \times \lceil T/\eta \rceil \times D_k} \to \mathbb{R}^{B \times T \times D_k}. \tag{10}$$

The final sequence is then given by

$$z_k = \Phi_\eta^{-1}(y^{(1)}, y^{(2)}, \ldots, y^{(\eta)}) \tag{11}$$

where $\Phi_\eta^{-1}(\Phi_\eta(x)) = x$ ensures that the rearrangement is bijective and preserves temporal fidelity. The reconstructed representation z_k is then passed to the subsequent block, enabling hierarchical refinement over progressively increasing dilation rates.

ii) *Scale-Aware Auxiliary Supervision*. To further enrich the temporal abstraction, we stack multiple **Temba blocks** with progressively increasing dilation rates, where the k-th block uses $\eta = k$. Each block refines the representation obtained from the previous one:

$$z_k = \text{Temba}_{\eta=k}(z_{k-1}); \quad k \geq 1 \tag{12}$$

enabling hierarchical modeling across temporal scales. Blocks with smaller dilation factors capture local and short-term variations, while those with larger dilation factors integrate long-range dependencies. Through this hierarchical stacking, the model gradually builds a comprehensive temporal understanding that spans multiple scales, enabling accurate and efficient reasoning over long untrimmed videos. To enhance discriminability and encourage scale-specific specialization, we employ a **scale-aware auxiliary supervision** mechanism. Each block is equipped with a lightweight classification head that produces block-level predictions $\hat{\text{Y}}^k$ from its output representation. The auxiliary loss for each block is defined as a standard Binary Cross Entropy (BCE) loss between its block-level predictions $\hat{\text{Y}}^k$ and the ground-truth labels Y:

$$\mathcal{L}_{\text{aux}} = \text{BCE}(\hat{\text{Y}}^k, \text{Y}) \tag{13}$$

ensuring that each block receives direct supervisory feedback. This auxiliary supervision constrains the temporal semantics of intermediate representations, maintaining each block's inductive bias toward its respective receptive field and dilation scale. By promoting such scale-aware discriminability, MS-Temba learns a hierarchy of temporally specialized features, thereby motivating the need for a dedicated *fusion module* to effectively integrate these multi-scale representations into a unified temporal encoding.

4.3. Multi-scale Mamba Fuser

The **Multi-scale Mamba Fuser (MS-Fuser)** aggregates the temporal representations at multiple temporal scales learned by the stacked Temba blocks to form a unified, scale-aware temporal embedding. Let $\{z_1, z_2, \ldots, z_K\}$ denote the outputs of the K Temba blocks, where each $z_k \in \mathbb{R}^{B \times T \times D_k}$ corresponds to the representation at dilation rate $\eta = k$. Since the feature dimensions D_k may vary, each z_k is first projected to a common embedding dimension E through a linear transformation with weight matrix $\mathbf{W}_k^f$ and bias $\boldsymbol{\beta}_k^f$ for block k. The projected features are then fused to obtain a unified representation z^f:

$$\tilde{z}_k = \mathbf{W}_k^f z_k + \boldsymbol{\beta}_k^f, \quad \tilde{z}_k \in \mathbb{R}^{B \times T \times E}. \tag{14}$$

$$z^f = \sum_{k=1}^{K} \tilde{z}_k \tag{15}$$

To capture cross-scale temporal dependencies, an additional SSM refines the fused representation through hidden state transitions:

$$h_t^f = \overline{\mathbf{A}^f} h_{t-1}^f + \overline{\mathbf{B}^f} z_t^f, \tag{16}$$

$$y_t^f = \mathbf{C}^f h_t^f. \tag{17}$$

The resulting sequence $\{y_t^f\}$ encodes both fine-grained and long-range temporal relationships, allowing MS-Fuser to effectively integrate complementary information from different dilation rates. This design ensures that earlier Temba blocks contribute detailed short-term cues, while later blocks provide coarse, long-horizon context, yielding a compact yet expressive multi-scale temporal representation well-suited for action detection in long untrimmed videos. Finally, the fused output y_t^f is passed through a classification head to obtain multiclass predictions $\hat{\text{Y}} = \{\hat{y}_t\}$ for each temporal segment for temporal action localization.

4.4. Training MS-Temba

For multi-label action detection, following prior works [11, 62], we optimize the network using a BCE loss:

$$\mathcal{L}_{\text{BCE}} = \text{BCE}(\hat{\text{Y}}, \text{Y}) \tag{18}$$

where Y denotes the ground-truth label, and $\hat{\text{Y}}$ is the model prediction. The overall training objective combines the primary BCE loss, the pairwise consistency loss and the averaged auxiliary losses:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \alpha \mathcal{L}_{\text{cons}} + \frac{\beta}{K} \sum_{k=1}^{K} \mathcal{L}_{\text{aux}}, \tag{19}$$

where α and β controls the strength of the consistency and auxiliary losses respectively. This composite objective enforces global discriminability, local temporal coherence, and scale-aware specialization, enabling robust temporal action detection in long untrimmed videos.

Table 1. **Comparison with the State-of-the-art**. **Bold** denotes best scores, underline denotes the second best.

Method	Visual Backbone	#Param (M)	TSU (mAP)	Charades (mAP)
Super-event [49]	I3D	26	17.2	18.6
TGM [50]	I3D	2	26.7	20.6
PDAN [10]	I3D	6	32.7	23.7
Coarse-Fine [33]	X3D	8	-	25.1
MLAD [62]	I3D	21	-	18.4
PointTAD [60]	I3D	-	-	22.1
HAAN [22]	I3D	-	-	25.2
CTRN [9]	I3D	11	33.5	25.3
MS-TCT [11]	I3D	87	33.7	**25.4**
DualDETR [76]	I3D	21	34.8	23.2
MS-Temba (ours)	I3D	17	**36.1**	**25.4**
TTM [53]	ViViT	89	-	26.3
AAN [12]	CLIP	-	41.3	32.0
MS-TCT [11]	CLIP	87	40.6	31.9
MS-Temba (ours)	CLIP	17	**44.0**	**33.6**

5. Experiments

Datasets. We evaluate MS-Temba, on two challenging, real-world, multi-label action detection datasets: Toyota Smarthome Untrimmed [8], and Charades [55]. The datasets contain densely labeled activities, making the task of action detection challenging. Toyota Smarthome Untrimmed (TSU) [8] is an untrimmed ADL video dataset capturing indoor activities related to daily living. It includes 51 action classes, with up to 5 co-occurring actions in a single frame, as well as composite actions. The average duration of the videos in TSU is 21 minutes, considerably higher than the other datasets making it challenging for action detection. Charades [55] is another large-scale dataset consisting of 9848 untrimmed videos spanning 157 actions. It contains 66500 annotations of daily activities with an average of 6.8 action instances per video. We evaluate our methods on these datasets using the '*action localization*' protocol described in [8], and [56], respectively, to detect frame-level actions and compare them with the temporal annotations.

Implementation Details. In MS-Temba, we use I3D [5] and CLIP-L/14 [51] as our visual backbones. Their feature dimensions D are 1024 and 768 respectively. Input features are initially transformed to an embedding of dimension $D_0 = 256$ before being passed to the first Temba block. MS-Temba consists of $K = 3$ Temba blocks with a feature expansion ratio of $\gamma = 1.5$. We follow the Mamba configurations of [74], with a state-dimension of 16. The dilation parameters are set to $\eta = k$ for Temba block k. For TSU, and Charades, the input Temporal dimensions are fixed at 2500, and 256, respectively. We set $\alpha = 100.0$ and $\beta = 1$ in all experiments. In the Supplementary material, we provide configuration details for each of the datasets. We use Adam optimizer [35] and cosine learning rate scheduler. All our models are trained on a single 24GB RTX A5000 GPU.

5.1. Comparison with State-of-the-Art

In Table 1, we benchmark MS-Temba against the state-of-the-art Temporal Action Detection methods on TSU and Charades datasets. The top section demonstrates the performance of models using an I3D visual backbone for feature extraction which is predominantly used in previous works [9–11, 50, 62]. In the TSU dataset, while other methods struggle to detect actions in the videos due to their longer duration compared to other datasets, MS-Temba outperforms the state-of-the-art methods, demonstrating its ability to learn effective long-range representations in the state space, directly tackling **Challenge 1**. We observe that MS-Temba achieves comparable performance with state-of-the-art [11] in Charades with $\mathbf{80}\%$ less parameters.

The bottom section of Table 1, presents the performance of temporal models with foundation models as the visual backbone. TTM [53] uses a ViViT [3] backbone, pretrained on the large-scale datasets JFT [59] and Kinetics-400 [34] and finetuned on the target dataset, for feature extraction whereas the other models use the pretrained foundation model CLIP [51]. We observe that MS-Temba trained using CLIP features achieves state-of-the-art performance on both datasets and even outperforms MS-Temba trained with I3D features on both TSU and Charades, highlighting the fact that foundation models learn more generalizable discriminative features compared to specialized models.

In Figure 5, we evaluate adaptive dilation in MS-Temba by comparing each Temba to its fixed-dilation counterpart. Temba Block 1 improves mAP on short actions (< 10 s), while Temba Block 3 excels on long actions (> 20 s). Thus, progressively increasing dilation across blocks enables MS-Temba to effectively capture both fine-grained short-term dynamics and long-range temporal dependencies in untrimmed videos aligning with the goal of **Challenge 2**.

Table 2. **Performance on Action Conditioned Metrics on TSU.** P_{AC} - Action Conditioned Precision; R_{AC} - Action Conditioned Recall; $F1_{AC}$ - Action Conditioned F1-score; mAP_{AC} - Action Conditioned Mean Average Precision. τ indicates the temporal window size.

Method	$\tau = 0$				$\tau = 20$				$\tau = 40$			
	$\mathbf{P}_{AC}$	$\mathbf{R}_{AC}$	$\mathbf{F1}_{AC}$	$\mathbf{mAP}_{AC}$	$\mathbf{P}_{AC}$	$\mathbf{R}_{AC}$	$\mathbf{F1}_{AC}$	$\mathbf{mAP}_{AC}$	$\mathbf{P}_{AC}$	$\mathbf{R}_{AC}$	$\mathbf{F1}_{AC}$	$\mathbf{mAP}_{AC}$
I3D [5]	12.84	3.73	5.79	13.94	24.10	6.14	9.79	21.39	25.05	6.12	9.84	20.41
I3D + MS-TCT [11]	23.26	18.15	20.40	25.74	34.19	22.67	27.26	37.12	35.49	23.86	28.53	36.35
I3D + MS-Temba (ours)	**25.68**	**19.90**	**22.43**	**27.03**	**38.13**	**24.59**	**29.90**	**38.86**	**38.96**	**24.84**	**30.34**	**38.05**
CLIP + MS-TCT [11]	26.49	21.34	23.64	31.14	40.78	30.93	35.18	43.44	42.66	32.40	36.80	43.36
CLIP + MS-Temba (ours)	**28.61**	**23.66**	**25.90**	**31.50**	**46.31**	**32.89**	**38.47**	**45.95**	**48.26**	**34.28**	**40.09**	**45.82**

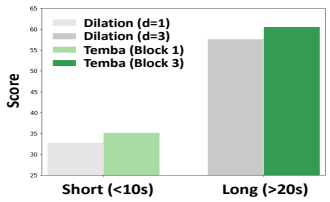


Figure 5. Impact of dilation in Temba for short and long actions.

Table 3. Impact of each component of MS-Temba.

Temporal Encoder	Dilated SSM	MS-Fuser	TSU (mAP)	Charades (mAP)
✗	✗	✗	24.7	22.9
✓	✗	✗	40.2	32.4
✓	✓	✗	42.5	32.5
✓	✓	✓	44.0	33.6

Table 4. Ablation on Loss components in MS-Temba.

$\mathcal{L}_{\text{cons}}$	$\mathcal{L}_{\text{aux}}$	TSU (mAP)	Charades (mAP)	Average (mAP)
✗	✗	42.1	33.6	41.9
✓	✗	42.1	33.7	41.9
✗	✓	43.0	33.1	42.8
✓	✓	44.0	33.6	43.8

Furthermore, in Table 2, we measure the MS-Temba's ability to model temporal dependencies and co-occurrence dependencies of actions on the TSU dataset, using the Action Conditioned Metrics introduced in [62]. The metrics compute Precision (P_{AC}), Recall (R_{AC}), F1-Score ($F1_{AC}$) and mAP (mAP_{AC}) which are conditioned on actions occurring within a temporal window τ. In Table 2, for $\tau = 0$ we observe that MS-Temba achieves state-of-the-art performance indicating its effectiveness in detecting densely distributed overlapping actions thus addressing **Challenge 3**. Furthermore, MS-Temba outperforms the baselines for $\tau > 0$ reaffirming MS-Temba's ability in capturing long-range dependencies between action instances.

Table 5. Localization Precision of MS-Temba at tIoU threholds.

Method	0.1	0.2	0.3	0.4	0.5	Avg
MS-TCT	42.8	40.6	36.8	31.5	28.5	36.1
MS-Temba	**49.1**	**46.4**	**41.6**	**34.2**	**31.0**	**40.5**

In Table 5, we further examine the localization performance of MS-Temba across varying tIoU thresholds. MS-Temba consistently improves over MS-TCT at every threshold, with the gains becoming more pronounced as the localization criterion becomes stricter, reflecting more accurate boundary alignment.

5.2. Ablation Study

Impact of each component in MS-Temba. Table 3 summarizes component-wise contributions. The baseline uses a simple classification head over CLIP features without temporal modeling. Adding a multi-layer Mamba temporal encoder [74] yields a large gain, highlighting the value of temporal modeling in long untrimmed videos. Replacing standard scanning with *Dilated Scanning* in SSMs improves TSU by $+2.3\%$ and maintains performance on Charades, capturing actions with varied durations. The MS-Fuser further boosts results by aggregating complementary cues across MS-Temba blocks. Together, these components deliver consistent, significant improvements over the CLIP baseline, showing the benefit of each module of MS-Temba.

Impact of Loss Components. Following previous work, the baseline model is trained using only the Binary Cross-Entropy ($\mathcal{L}_{\text{BCE}}$) classification loss for action detection. As shown in Table 4, incorporating the proposed consistency loss ($\mathcal{L}_{\text{cons}}$) alone yields minimal improvement, as it primarily acts as a regularizer on the projection space of SSM outputs (C_i of i-th SSM in a block) without directly influencing the final prediction confidence. Additionally, introducing the scale-aware auxiliary loss ($\mathcal{L}_{\text{aux}}$) leads to a clear performance gain by encouraging intermediate Temba blocks to learn discriminative, scale-specific temporal representations. The best performance is achieved when both losses are combined with $\mathcal{L}_{\text{BCE}}$, providing a $+1.9\%$ increase in average mAP across datasets. The proposed combination of losses is therefore instrumental in promoting coherent temporal alignment across dilated SSMs and enhancing multi-scale representation learning within MS-Temba.

Table 6. Impact of alignment objectives among dilated SSMs in Temba blocks.

Loss	TSU	Ch
No Loss	43.0	33.1
Correlation Loss	43.5	33.1
Consistency Loss	44.0	33.6

Table 7. Analysis of multi-scale fusion strategies.

Method	TSU	Ch
Temba Block 3	41.8	32.2
Concat + Proj.	42.2	32.8
Concat + Proj. + SSM	42.4	32.4
Sum + Proj.	42.5	32.5
Sum + Proj. + SSM	44.0	33.6

5.3. Design Choices

Choice of Projection Alignment in Temba Blocks. Table 6 ablates coherence among dilated SSMs within each Temba block ($\eta \geq 2$). The *No Loss* baseline trains SSMs independently, without any constraint on their projected subspaces. Adding a correlation regularizer that encourages

linear co-variation between projection matrices $\mathbf{C}_i$ and $\mathbf{C}_j$ yields a modest gain. Our *consistency loss*, which maximizes pairwise cosine similarity between *normalized* projection matrices, delivers the best and most stable improvements on both datasets. This indicates that aligning projection subspaces across parallel SSMs promotes coherent temporal semantics while preserving complementary scale-specific cues for temporal action detection.

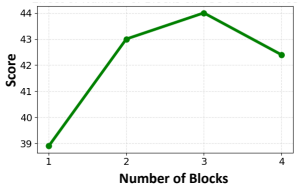


Figure 6. Analysis of number of Temba blocks.

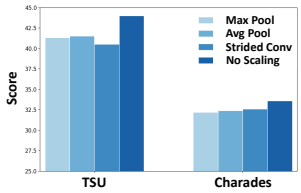


Figure 7. Impact of pooling temporal dimension.

Number of Temba Blocks. As shown in Figure 6, increasing the number of stacked Temba blocks from one to three steadily improves performance, indicating that progressively larger dilations help capture richer temporal hierarchies. However, adding a fourth block causes a drop in accuracy, suggesting diminishing returns from deeper stacking. A three-block configuration achieves the best trade-off between temporal coverage and model effectiveness.

Impact of scaling temporal dimension in Temba Blocks. Prior work handles untrimmed videos at multiple temporal scales by first downsampling the temporal dimension and later interpolating them back to the input resolution. We evaluate analogous temporal scaling strategies within each Temba block. As shown in Figure 7, max/avg pooling markedly degrades performance, indicating loss of fine-grained cues; strided convolutions also underperform due to temporal discontinuities. MS-Temba performs best with no explicit scaling, allowing each block to operate at native resolution, preserving subtle transitions while capturing long-range dependencies in untrimmed videos.

Analysis of fusion strategies. Table 7 compares multi-scale fusion in the MS-Fuser. Concatenating block outputs with a projection modestly outperforms the final Temba output, and adding an SSM on top yields limited extra benefit. In contrast, summation-based fusion better aligns features across scales, and an additional SSM further strengthens long-range temporal modeling. This combination attains the highest performance on both datasets, indicating that additive fusion preserves complementary cross-block information while the SSM refines the fused representation through temporal integration.

6. Beyond TAD: Video Summarization

In addition to TAD, we demonstrate the generality of MS-Temba on the task of video summarization (Table 8). This task condenses untrimmed videos into concise yet comprehensive summaries of key events and requires reasoning over both short-term visual changes and long-range temporal dependencies to estimate frame-level importance.

Table 8. Performance of MS-Temba on Video Summarization.

	SumMe		TVSum	
Method	τ	ρ	τ	ρ
Random	0.000	0.000	0.000	0.000
Human	0.205	0.213	0.177	0.204
Visual only				
VASNet [17]	0.160	0.170	0.160	0.170
DMASum [65]	0.063	0.089	0.203	0.267
RR-STG [75]	0.211	0.234	0.162	0.212
CSTA[57]	0.246	0.274	0.194	0.255
Visual + Text				
CLIP-It [48]	–	–	0.108	0.147
A2Summ [29]	0.108	0.129	0.137	0.165
MSVA [23]	0.200	0.230	0.190	0.210
SSPVS [37]	0.192	0.257	0.181	0.238
LLMVS [36]	0.253	0.282	0.211	0.275
CLIP (w/o MS-Temba)	0.197	0.221	0.207	0.270
MS-Temba	**0.254**	**0.284**	**0.221**	**0.289**

Following Clip-It [48], we generate frame-level descriptions for uniformly sampled frames and extract CLIP features [51]. We form video features by stacking visual CLIP embeddings and apply cross-attention with the text descriptors; the cross-attended sequence is then processed by MS-Temba. The classification head is replaced with a linear regression head [57] to predict per-frame importance.

We evaluate the performance of the cross-attended input features (w/o MS-Temba) and MS-Temba on Video Summarization on the TVSum [58] and SumMe [28] benchmarks (details in Supplementary). On both the datasets, MS-Temba attains state-of-the-art results. We attribute the gains to the dilated SSMs within each Temba, which enable modeling of temporal dependencies over varying time horizons, capturing both fine-grained transitions and long-range semantic structures essential for effective summarization. These findings reinforce the versatility of MS-Temba as a generic framework for long-video understanding.

7. Conclusion

This work presents MS-Temba, a multi-scale Mamba-based approach for TAD in Untrimmed ADL Videos. By integrating dilation into the SSM formulation, and guiding each temporal scale through alignment and targeted supervision, MS-Temba learns short-term atomic actions as well as extended activities. The Multi-scale Mamba Fuser learns a robust unified representation from scale-specific representations, enabling accurate localization in densely annotated ADL settings. MS-Temba achieves state-of-the-art performance on TSU and Charades setting a new benchmark for TAD in Activities of Daily Living with only 17M parameters, significantly outperforming existing methods. Beyond action detection, MS-Temba also generalizes to Video Summarization, achieving state-of-the-art in standard benchmarks. This establishes MS-Temba as a powerful and versatile framework for long-video understanding.

Acknowledgements

We would like to thank Manish Kumar Govind and Ipsita Ghosh for their inputs on the experiments on Video Summarization. This research is supported in part by the National Science Foundation (IIS-2245652) and the UNC Charlotte AI4Health Center Ignite Grant.

References

[1] S. Mohsen Amiri, Mahsa T. Pourazad, Panos Nasiopoulos, and Victor C.M. Leung. Non-intrusive human activity monitoring in a smart home environment. In *2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013)*, pages 606–610, 2013. 1

[2] Anwai Archit and Constantin Pape. Vim-unet: Vision mamba for biomedical segmentation. *arXiv preprint arXiv:2404.07705*, 2024. 3

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, 2021. 6

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 2

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Conference on Computer Vision and Pattern Recognition*, pages 4724–4733. IEEE, 2017. 4, 6, 7

[6] Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Zhe Chen, Zhiqi Li, Jiahao Wang, Kunchang Li, Tong Lu, and Limin Wang. Video mamba suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*, 2024. 2, 3

[7] Xu Chen, Yang Li, Yahong Han, and Jialie Shen. Ex pede herculem, predicting global actionness curve from local clips. In *Proceedings of the 33rd ACM International Conference on Multimedia*, page 7239–7247, New York, NY, USA, 2025. Association for Computing Machinery. 2

[8] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota Smarthome Untrimmed: Real-World Untrimmed Videos for Activity Detection. *arXiv preprint arXiv:2010.14982*, 2020. 6

[9] Rui Dai, Srijan Das, and Francois F Bremond. CTRN: Class Temporal Relational Network For Action Detection. In *BMVC 2021 - The British Machine Vision Conference*, Virtual, United Kingdom, 2021. 6

[10] Rui Dai, Srijan Das, Luca Minciullo, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. PDAN: Pyramid Dilated Attention Network for Action Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2970–2979, 2021. 2, 4, 6

[11] Rui Dai, Srijan Das, Kumara Kahatapitiya, Michael Ryoo, and Francois Bremond. MS-TCT: Multi-Scale Temporal ConvTransformer for Action Detection. In *CVPR*, 2022. 2, 3, 4, 6, 7

[12] Rui Dai, Srijan Das, Michael Ryoo, and Francois Bremond. AAN : Attributes-Aware Network for Temporal Action Detection. In *BMVC 2023 - The 34th British Machine Vision Conference*, Aberdeen, United Kingdom, 2023. 6

[13] Xiyang Dai, Bharat Singh, Joe Yue-Hei Ng, and Larry Davis. Tan: Temporal aggregation network for dense multi-label action recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 151–160. IEEE, 2019. 2

[14] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[15] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *International Conference on Computer Vision*, 2019. 1

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021. 3

[17] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. In *Asian conference on computer vision*, pages 39–54. Springer, 2018. 8

[18] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 3

[19] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3575–3584, 2019. 2, 4

[20] Alireza Fathi, Xiaofeng Ren, and James M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288, 2011. 3

[21] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4768–4777, 2017. 1

[22] Zikai Gao, Peng Qiao, and Yong Dou. Haan: Human action aware network for multi-label temporal action detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 5059–5069, New York, NY, USA, 2023. Association for Computing Machinery. 3, 6

[23] Junaid Ahmed Ghauri, Sherzod Hakimov, and Ralph Ewerth. Supervised video summarization via multiple feature sets with parallel attention. *arXiv preprint arXiv:2104.11530*, 2021. 8

[24] Joshua Gleason, Rajeev Ranjan, Steven Schwarcz, Carlos Castillo, Jun-Cheng Chen, and Rama Chellappa. A

proposal-based solution to spatio-temporal action detection in untrimmed videos. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 141–150. IEEE, 2019. 2

[25] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2, 3

[26] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 2, 3

[27] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *European conference on computer vision*, pages 222–241. Springer, 2024. 3

[28] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European Conference on Computer Vision*, 2014. 8

[29] Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14867–14878, 2023. 8

[30] Xuanhua He, Ke Cao, Jie Zhang, Keyu Yan, Yingying Wang, Rui Li, Chengjun Xie, Danfeng Hong, and Man Zhou. Pan-mamba: Effective pan-sharpening with state space model. *Information Fusion*, 115:102779, 2025. 3

[31] Md Mohaiminul Islam, Mahmudul Hasan, Kishan Shamsundar Athrey, Tony Braskich, and Gedas Bertasius. Efficient movie scene detection using state-space transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18749–18758, 2023. 3

[32] Yu-Gang. Jiang, Jingen Liu, Amir Roshan Zamir, George Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. THUMOS Challenge: Action Recognition with a Large Number of Classes. http://crcv.ucf.edu/THUMOS14/, 2014. 2

[33] Kumara Kahatapitiya and Michael S Ryoo. Coarse-Fine Networks for Temporal Activity Detection in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8385–8394, 2021. 6

[34] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 6

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6

[36] Min Jung Lee, Dayoung Gong, and Minsu Cho. Video summarization with large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18981–18991, 2025. 8

[37] Haopeng Li, Qiuhong Ke, Mingming Gong, and Tom Drummond. Progressive video summarization via multimodal self-supervised learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5584–5593, 2023. 8

[38] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pages 237–255. Springer, 2024. 2, 3

[39] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 988–996. ACM, 2017. 2

[40] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2

[41] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3889–3898, 2019. 2

[42] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Cheng Li, Yong Liang, Guangming Shi, Yizhou Yu, Shaoting Zhang, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 615–625. Springer, 2024. 3

[43] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024. 3

[44] Yunqing Liu, Nan Zhang, Fangjun Wang, Kengo Murata, Takuma Yamamoto, Osafumi Nakayama, Genta Suzuki, and Zhiming Tan. Mamba only glances once (mogo): A lightweight framework for efficient video action detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3

[45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*, 2021. 3

[46] Hui Lu, Albert A. Salah, and Ronald Poppe. Snakes and ladders: Two steps up for videomamba. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24234–24244, 2025. 2, 3

[47] Chaoqi Luo, Yiping Xie, and Zitong Yu. Physmamba: Efficient remote physiological measurement with slowfast temporal difference mamba. In *Chinese Conference on Biometric Recognition*, pages 248–259. Springer, 2024. 3

[48] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *Advances in neural information processing systems*, 34:13988–14000, 2021. 8

[49] AJ Piergiovanni and Michael S Ryoo. Learning latent super-events to detect multiple activities in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 6

[50] AJ Piergiovanni and Michael S Ryoo. Temporal gaussian mixture layer for videos. *International Conference on Machine Learning (ICML)*, 2019. 2, 6

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 4, 6, 8

[52] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Finetuned clip models are efficient video learners. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[53] Michael S Ryoo, Keerthana Gopalakrishnan, Kumara Kahatapitiya, Ted Xiao, Kanishka Rao, Austin Stone, Yao Lu, Julian Ibarz, and Anurag Arnab. Token turing machines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19070–19081, 2023. 3, 6

[54] Jiayi Shao, Xiaohan Wang, Ruijie Quan, Junjun Zheng, Jiang Yang, and Yi Yang. Action sensitivity learning for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13457–13469, 2023. 3

[55] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *European Conference on Computer Vision(ECCV)*, 2016. 6

[56] Gunnar A Sigurdsson, Santosh Divvala, Ali Farhadi, and Abhinav Gupta. Asynchronous temporal fields for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 585–594, 2017. 6

[57] Jaewon Son, Jaehun Park, and Kwangsu Kim. Csta: Cnn-based spatiotemporal attention for video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18847–18856, 2024. 8

[58] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5179–5187, 2015. 8

[59] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, pages 843–852. IEEE Computer Society, 2017. 6

[60] Jing Tan, Xiaotong Zhao, Xintian Shi, Bin Kang, and Limin Wang. Pointtad: Multi-label temporal action detection with learnable query points. In *Advances in Neural Information Processing Systems*, pages 15268–15280. Curran Associates, Inc., 2022. 2, 3, 6

[61] Haoran Tang, Meng Cao, Jinfa Huang, Ruyang Liu, Peng Jin, Ge Li, and Xiaodan Liang. Muse: Mamba is efficient multi-scale learner for text-video retrieval. *arXiv preprint arXiv:2408.10575*, 2024. 3

[62] Praveen Tirupattur, Kevin Duarte, Yogesh Rawat, and Mubarak Shah. Modeling multi-label action dependencies for temporal action localization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3, 6, 7

[63] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 3

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[65] Junyan Wang, Yang Bai, Yang Long, Bingzhang Hu, Zhenhua Chai, Yu Guan, and Xiaolin Wei. Query twice: Dual mixture attention meta learning for video summarization. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 4023–4031. ACM, 2020. 8

[66] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 578–588. Springer, 2024. 3

[67] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE international conference on computer vision*, pages 5783–5792, 2017. 2

[68] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-TAD: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10156–10165, 2020. 2

[69] Yijun Yang, Zhaohu Xing, and Lei Zhu. Vivim: a video vision mamba for medical video object segmentation. *arXiv preprint arXiv:2401.14168*, 2024. 3

[70] Olga Zatsarynna, Emad Bahrami, Yazan Abu Farha, Gianpiero Francesca, and Juergen Gall. Manta: Diffusion mamba for efficient and effective stochastic long-term dense action anticipation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 3

[71] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[72] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. HACS: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8668–8678, 2019. 2, 3

[73] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017. 2

[74] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Forty-first International Conference on Machine Learning*, 2024. 2, 3, 6, 7

[75] Wencheng Zhu, Yucheng Han, Jiwen Lu, and Jie Zhou. Relational reasoning over spatial-temporal graphs for video summarization. *IEEE Transactions on Image Processing*, 31: 3017–3031, 2022. 8

[76] Yuhan Zhu, Guozhen Zhang, Jing Tan, Gangshan Wu, and Limin Wang. Dual detrs for multi-label temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18559–18569, 2024. 6

[77] Nikola Zubic, Mathias Gehrig, and Davide Scaramuzza. State space models for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5819–5828, 2024. 3