electronics

Article

ETMamba: An Effective Temporal Model for Video Action Recognition

Rundong Hong [1], Changji Wen [1,2,*], Patrick Sun [3], Leyao Zhang [1], Zhuozhen Niu [1], Yaqi Shi [1], Chenshuang Li [1], Mingqi Li [4], Hengqiang Su [1] and Hongbing Chen [1,*]

[1] College of Information and Technology, Jilin Agricultural University, Changchun 130118, China; 20231212@mails.jlau.edu.cn (R.H.); 20241247@mails.jlau.edu.cn (L.Z.); 20241293@mails.jlau.edu.cn (Z.N.); 20241306@mails.jlau.edu.cn (Y.S.); lcs@mails.jlau.edu.cn (C.L.); suhengqiang@jlau.edu.cn (H.S.)
[2] Key Laboratory of Modern Agricultural Equipment and Technology (Jiangsu University), Ministry of Education, Zhenjiang 212001, China
[3] Letter and Science, University of Wisconsin, Madison, WI 53707, USA; psun4782@gmail.com
[4] Midea Group (Shanghai) Co., Ltd., Shanghai 201700, China; limq155@midea.com
* Correspondence: changjiw@jlau.edu.cn (C.W.); chenhongbing@jlau.edu.cn (H.C.)

Abstract

Video action recognition faces persistent challenges in balancing accuracy with computational efficiency. While state space models, such as Mamba, have emerged with linear complexity advantages, they exhibit inefficiency in capturing critical spatiotemporal dependencies within video data. To address this core limitation, this paper proposes ETMamba, an enhanced architecture built upon the Mamba baseline. The ETMamba achieve performance breakthroughs via three core innovation modules: (1) the Spatiotemporal Feature Preservation module retains complete original spatiotemporal correlations before data flattening, solving the problem of spatiotemporal feature loss; (2) the Efficient Bidirectional Sharing strategy accurately models bidirectional temporal dependencies, enhancing key temporal dynamic information; and (3) the Spatiotemporal Collaborative Modulation mechanism combines global temporal and local spatial information to achieve collaborative capture of long-short term dependencies and fine-grained features. We conduct experiments on multiple benchmark datasets, achieving recognition accuracies of 88.3%, 74.6%, 75.7%, and 98.1% on Kinetics-400, Something-Something V2, HMDB-51, and Breakfast datasets, respectively, while maintaining low to medium computational complexity.

Keywords: video action recognition; state space models; Mamba; spatiotemporal feature fusion; multi-scale feature enhancement

Check for updates

Academic Editor: Stefanos Kollias

Received: 11 February 2026
Revised: 19 March 2026
Accepted: 19 March 2026
Published: 23 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Video behavior recognition is a fundamental research topic in computer vision, which is based on the effective modeling of spatiotemporal dynamic features in video sequences. The visual content of the same moment in the video is affected by the temporal correlation between the preceding and following frames, while the action subject within a single frame is tightly coupled with the local spatial region. Therefore, the key challenge in video behavior recognition is how to efficiently model long-range dependencies while accurately capturing fine-grained local spatial information. For decades, research in video behavior recognition has primarily focused on two core architectures: convolutional neural networks (CNNs) and Transformers. However, both approaches face inherent limitations in balancing

accuracy and efficiency. As shown in Figure 1, we compared the effective receptive field (ERF) [1] distributions across different architectures.

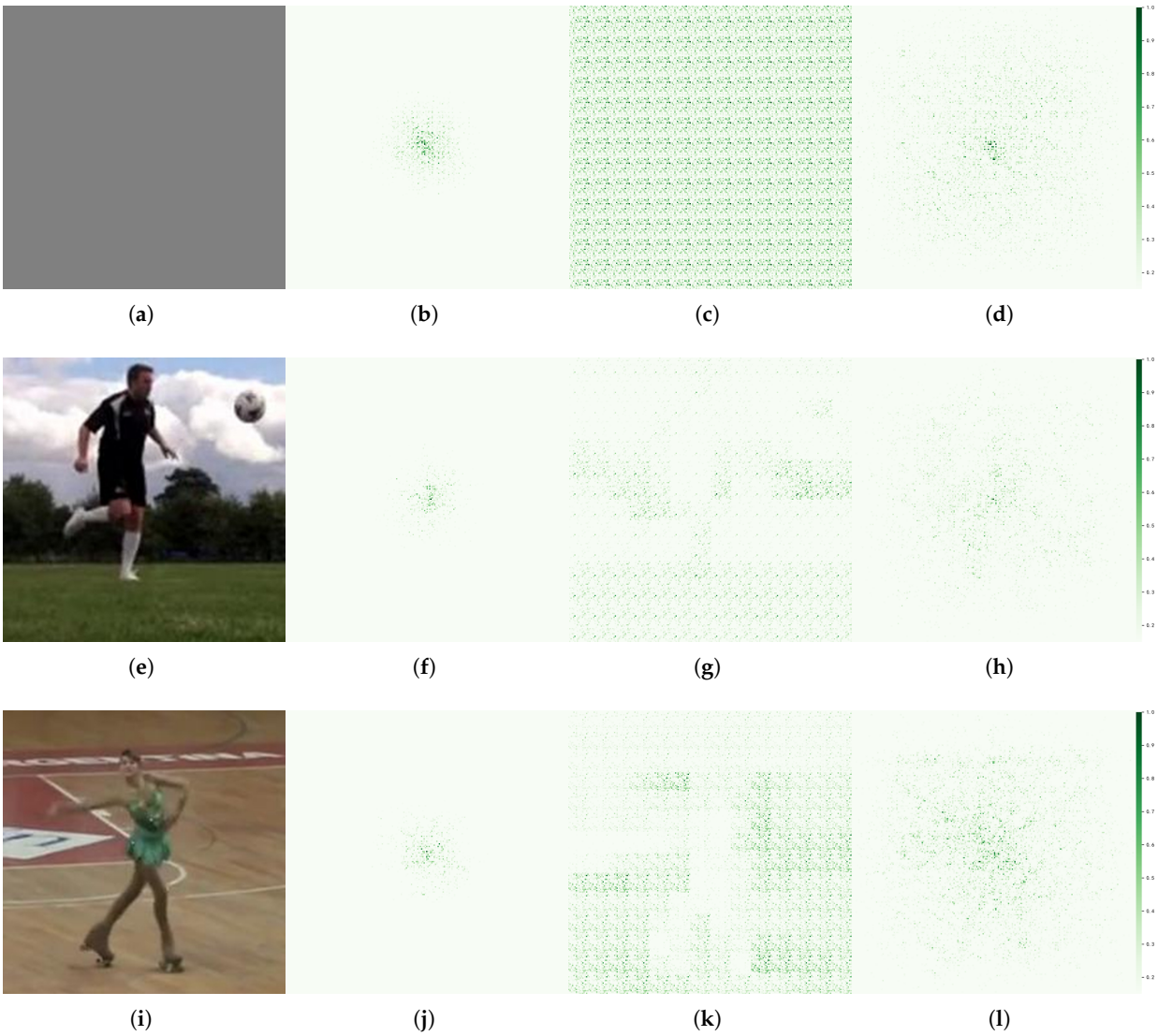


Figure 1. Visualizations of ERF for different architectures and the proposed ETMamba. The more widely the dark area is distributed, the larger the effective receptive field is. (**a**,**e**,**i**) Original image. (**b**,**f**,**j**) CNNs based. (**c**,**g**,**k**) Transformer based. (**d**,**h**,**l**) ETMamba(Ours).

CNN-based methods, by leveraging parallel convolution operations, can reduce local spatiotemporal redundancy to some extent. However, constrained by the receptive field of convolution kernels, their ERF are highly concentrated in central regions, making it difficult to model long-distance behavioral logic across multiple frames. The Transformer-based approach achieves global ERF through self-attention mechanisms, effectively modeling long-range dependencies. However, its computational complexity increases quadratically with video sequence length due to the growing number of visual tokens, resulting in higher inference and training costs. To address these challenges, state space models (SSMs), with Mamba [2] as a prime example, provide an ideal solution. Mamba demonstrates linear scalability in sequence length, and its selective scanning operation (S6) significantly reduces training and inference costs while avoiding the computational complexity drawbacks of Transformers. Meanwhile, its robust representation capability efficiently captures the global dependencies in video data, effectively addressing the limitation of CNNs in identifying such relationships. Mamba has demonstrated promising performance in tasks such as

image recognition and object detection, providing a viable technical approach for efficient video behavior recognition. In this context, studies like VideoMamba [3] have preliminarily validated Mamba's effectiveness in video tasks. However, as it lacks dedicated design for video spatiotemporal features, its spatiotemporal modeling capability still has significant room for improvement, resulting in a precision gap compared to advanced Transformer models.

This study systematically evaluates the feature modeling capabilities of CNNs, Transformers, and Mamba architectures in video behavior recognition. The analysis reveals two critical limitations of the traditional Mamba architecture when processing video data: First, it tends to disrupt the original spatiotemporal structure and cause information loss during dimensionality transformation and sequence flattening. Second, it struggles to effectively model global temporal dependencies and local spatial features in a coordinated manner, resulting in an inability to achieve optimal balance between accuracy and efficiency. To address these challenges, this study introduces structural enhancements to VideoMamba, resulting in the innovative ETMamba model. Building upon the linear-complexity Mamba architecture, ETMamba employs multi-module collaborative design to expand effective receptive fields, thereby achieving more robust spatiotemporal feature modeling. Specifically, we first enhance the spatiotemporal structure preservation before sequence flattening to reduce the loss of adjacent information. The dual-branch architecture is adopted to realize the joint modeling of global temporal and local spatial features through bidirectional temporal branching and multi-scale spatial enhancement branching, which achieves the collaborative capture of directional features and fine-grained features. The main contributions of this paper are as follows:

(1) Spatiotemporal Feature Preservation (STFP) module: We propose a module that preserves spatiotemporal correlations before feature flattening. It combines depthwise separable convolutions and gated fusion to retain spatial and temporal dependencies, alleviating information loss in Mamba-based video models and supplying better features for spatiotemporal modeling.

(2) Efficient Bidirectional Sharing (EBS) Strategy: We propose a bidirectional temporal modeling mechanism for Mamba. By sharing core SSM parameters and separating forward and backward branches, it accurately models video bidirectional temporal dependencies. A directional bias mechanism is adopted to enhance critical temporal information, overcoming the unidirectional limitation of Mamba.

(3) The Spatiotemporal Collaborative Modulation (STCM) mechanism: We propose a dual-branch architecture that embeds multi-scale dilated convolutions into EBS, unifying global temporal modeling and local spatial feature extraction. This mechanism achieves the synergistic optimization of long- and short-range dependencies with fine-grained features, a capability not achieved by existing Mamba-based video recognition methods.

(4) Outstanding performance: Extensive experiments on four benchmark datasets (Kinetics-400 [4], Something-Something V2 [5], HMDB-51 [6], and Breakfast [7]) show that our method achieves an excellent balance between recognition accuracy and computational complexity, with Top-1 accuracies of 88.3%, 74.6%, 75.7%, and 98.1%, respectively. The results provide novel technical insights and practical references for the field of video action recognition.

2. Related Work

2.1. Traditional Action Recognition Methods

Traditional video action recognition methods mainly revolve around two architectures: CNNs and Transformers. Their core goal is to capture spatiotemporal dependencies in videos, but both methods face insurmountable bottlenecks due to their inherent charac-

teristics. Methods based on CNNs have become the mainstream solution for early video action recognition by virtue of their local inductive bias. These methods adapt to the spatiotemporal characteristics of video data by extending the dimensions and structures of convolution operations. For example, TSM [8] introduces 2D CNNs on the time channel to achieve the efficient exchange of inter-frame information. AFNet [9] dynamically selects salient frames using 2D CNNs to reduce redundant computations. SlowFast [10] constructs a dual-branch path focusing on temporal feature extraction and spatial feature extraction respectively. CSTANet [11] further explores the fusion correlations between temporal and spatial features within convolution channels on the basis of the dual-branch framework. ST-MFO [12] optimizes feature expression by suppressing background noise interference in the spatiotemporal branch framework. However, the receptive field of CNNs is limited by the size of convolution kernels, making it difficult to effectively model the behavioral logical correlations across multiple frames in video sequences, which becomes a core bottleneck for performance improvement. In contrast, Transformer-based methods break through the limitation of convolution receptive fields through self-attention mechanisms, realizing direct modeling of global dependencies. For example, TimeSformer [13] and VidTr [14] propose separable attention mechanisms by splitting spatiotemporal self-attention to complete video classification while ensuring modeling capability. Swin Transformer [15] restricts self-attention computation within local windows, enhancing local feature capture while reducing overhead. MViTv2 [16] adopts a design strategy of increasing channel dimensions and synchronously reducing spatial resolution to achieve optimization of computational costs. FSformer [17] learns short-term behavioral differences by constructing a dual-branch architecture. UniFormer [18] innovatively introduces 3D CNNs to fuse the advantages of CNNs and Transformers, and its subsequent improved version UniFormerV2 [19] further combines Multi-Head Relation Aggregator (MHRA) to enhance feature correlation modeling capability. Such methods can effectively capture the long-term motion trajectories of objects and cross-frame interactions in videos, but they need to compute the similarity between all tokens, leading to a quadratic increase in computational complexity, resulting in high computational overhead and memory consumption for the model during training and inference phases.

2.2. Mamba-Based Methods

The bottlenecks exposed by traditional CNNs and Transformer methods in long-sequence processing have promoted the research and development of models with low complexity and high representation capability. SSMs have become an important technical breakthrough direction in the field of video action recognition due to their linear computational complexity and efficient long-range dependency modeling capability. Based on dynamic system theory, SSMs model sequence data through state transition equations, which can accurately capture dependencies in long sequences under the premise of strictly controlling computational and memory overhead, providing a new technical path for solving the core pain points of traditional methods. As a representative model of SSMs, Mamba significantly improves computational efficiency by introducing gating multi-layer perceptrons and Selective Scan technology (S6) into the architecture. Inspired by this, researchers have begun to focus on the Mamba architecture. VideoMamba first applied Mamba to the video field, verifying its feasibility in action recognition. Subsequently, VideoMambaPro [20] proposed mask reverse computation and element-wise residual connection strategies. StableMamba [21] solved the scalability problem of SSMs through a Mamba-Attention interleaved architecture. VideoMAP [22] adopted a 4:1 ratio Mamba-Transformer hybrid framework and combined frame-wise masked autoregressive pre-training strategy to improve sample efficiency. MambaVL [23] optimized data pro-

cessing performance by sharing state transition matrices through a selective state space modal fusion mechanism. Mamba-ND [24] designed row-major orderings to alternately process input data of different dimensions. Although existing Mamba-based methods have achieved remarkable progress in video understanding, there remains a considerable accuracy gap compared with state-of-the-art Transformer models, leaving ample room for further optimization. We argue that the core bottleneck lies in Mamba's insufficient capability for spatiotemporal relationship modeling. Unlike Transformer, which is equipped with a self-attention mechanism featuring inherent global correlation capability, Mamba was originally designed for 1D sequence modeling. When applied to video processing tasks, it mandatorily requires flattening high-dimensional spatiotemporal input into 1D sequences to complete core computation. Its core serpentine sequential scanning strategy can only capture the correlation between consecutive tokens in the flattened 1D sequence. The dual operations of input flattening and fixed sequential scanning inevitably destroy the native spatiotemporal structure of video data, resulting in irreversible information loss, a limitation that has been widely validated and discussed in recent studies [25,26]. To address these challenges, this paper proposes the ETMamba model based on VideoMamba, with targeted optimization of the Mamba architecture through three core modules: the Spatiotemporal Feature Preservation (STFP) module, the Efficient Bidirectional Sharing (EBS) strategy, and the Spatiotemporal Collaborative Modulation (STCM) mechanism. The proposed model aims to achieve a favorable balance between low computational overhead and high recognition accuracy, and effectively resolve the core pain points of existing Mamba-based video understanding methods.

3. Method

3.1. Overall Architecture

ETMamba is a basic Mamba-style model, but it is elaborately designed to retain the spatiotemporal information correlations of video data and capture complex global temporal features and local spatial features. Specifically, the model constructs the overall framework through three key modules: Spatiotemporal Feature Preservation (STFP), Efficient Bidirectional Sharing (EBS), and Spatiotemporal Collaborative Modulation (STCM). The core design concept of this framework is to optimize the extraction and modeling effect of video spatiotemporal features through the collaborative work of each module. The network structure of ETMamba is shown in Figure 2.

For simplicity, we take a video sequence with T frames as an example. Video data is first input into the Spatiotemporal Feature Preservation (STFP) module, which extracts and fuses original spatiotemporal features through depthwise separable convolution and a gating mechanism, retaining the spatial position and temporal correlation information of high-dimensional data. Then, a 3D convolution of size $1 \times 16 \times 16$ is used to project the input video sequence $X \in \mathbb{R}^{T \times H \times W \times C}$ into L non-overlapping spatiotemporal patches $X_p \in \mathbb{R}^{L \times C}$ for subsequent position encoding and category encoding. Among them, H and W represent the height and width of the input video frames respectively, C is the number of channels, T is the number of frames of the input video sequence, and $L = t \times h \times w (t = T, h = H/16, w = H/16)$. The encoding process of input sequence X can be expressed as

$$X = [X_{\text{cls}}, X] + p_{\text{s}} + p_{\text{t}}, \tag{1}$$

where X_{cls} is a learnable classification token concatenated at the beginning of the sequence. Since state space model modeling is sensitive to token positions, we follow the design idea of VideoMamba and add a learnable spatial position embedding $p_{\text{s}} \in \mathbb{R}^{(h \times w+1) \times C}$ and an additional temporal position embedding $p_{\text{t}} \in \mathbb{R}^{t \times C}$ to retain spatiotemporal position information.

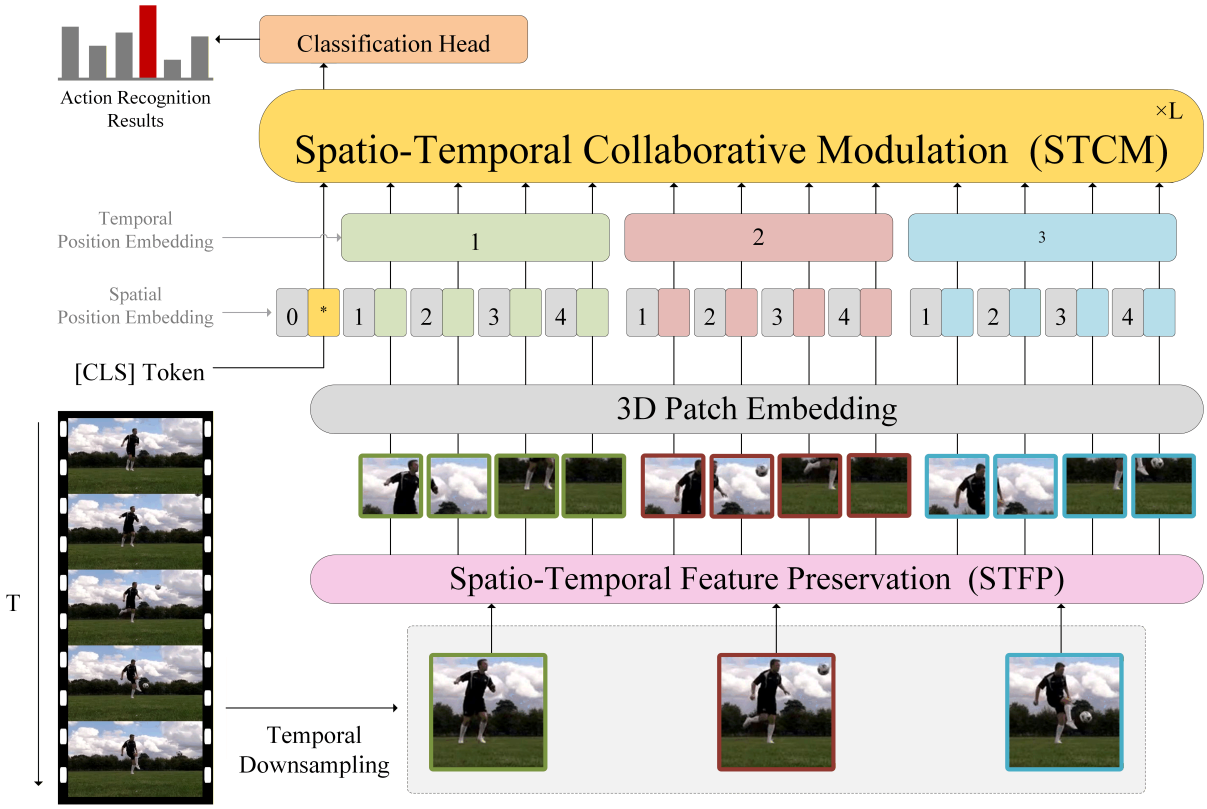


Figure 2. The overall architecture of ETMamba. Here, '*' denotes the [CLS] token.

Subsequently, the encoded sequence is fed into the Spatiotemporal Collaborative Modulation (STCM) Mamba Block stacked in L layers. Each STCM consists of two core components: one is the global temporal modeling module based on the Efficient Bidirectional Sharing (EBS) strategy, which splits the input into forward and backward branches and shares SSMs parameters to accurately model bidirectional temporal dependencies; the other is the local spatial modeling module based on multi-scale dilated convolution, which captures local spatial details at different fine-grained levels. Finally, the integrated high-dimensional features are fed into a classification head to determine the final action category, thereby achieving precise recognition of action behaviors in videos.

3.2. Spatiotemporal Feature Preservation (STFP)

Mamba was originally designed for natural language processing tasks and has inherent limitations when applied to human action recognition tasks based on video data: videos are three-dimensional data that need to be flattened to adapt to the model, which will inevitably lose spatial information during the dimensionality reduction process. Although position encoding can partially restore the spatial relationships between blocks during context modeling, the restored information is sparse and difficult to capture accurately, leading to the loss of the model's original spatial features and insufficient attention to the boundary areas of blocks. To address this problem, we propose to perform feature extraction before block division, specifically using depthwise separable convolution (DWConv) and a gating mechanism to achieve this task. To completely retain the original information, we combine the original features with deep features through a parameter-free gating strategy and further strengthen the fusion effect of the two through convolution. The network structure of STFP is shown in Figure 3.

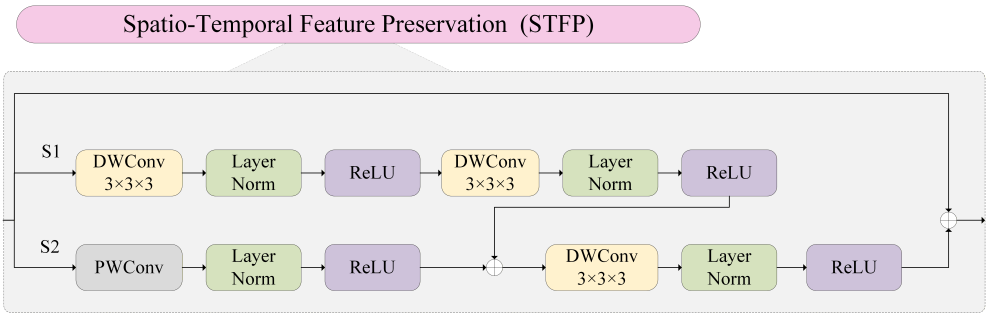


Figure 3. The Spatiotemporal Feature Preservation (STFP) module.

Specifically, the network of STFP includes two branch structures. For the input sequence X, the upper branch S_1 consists of two $3 \times 3 \times 3$ depthwise separable convolutions, each followed by normalization and ReLU activation function. The purpose of this structure is to capture local spatial context and neighborhood information through multiple convolution operations, which is consistent with our basic goal of extracting spatial information. The lower branch S_2 first performs a linear transformation on the channel dimension of the input data through a $1 \times 1 \times 1$ pointwise convolution while retaining the spatial information of the original data. In this way, the original input can be directly transmitted to subsequent network layers. Then, it is added pixel-wise with the output result of branch S_1, and features are fused through a $3 \times 3 \times 3$ depthwise separable convolution to achieve deeper feature integration. Finally, the output of branch S_2 is residual connected with the original input features, which can not only retain the direct block information of the original input but also combine it with the complex spatiotemporal block features extracted by the branches. STFP can be expressed as follows.

$$S_1 = \mathrm{ReLU}\big(\mathrm{LN}\big(\mathrm{DW}_3\big(\mathrm{ReLU}\big(\mathrm{LN}\big(\mathrm{DW}_3(X)\big)\big)\big)\big)\big), \tag{2}$$

$$S_2 = \mathrm{ReLU}\big(\mathrm{LN}\big(\mathrm{DW}_3\big(S_1 + \mathrm{ReLU}\big(\mathrm{LN}\big(\mathrm{PW}(X)\big)\big)\big)\big)\big), \tag{3}$$

$$\mathrm{STFP} = S_2 + X, \tag{4}$$

where S_1 and S_2 represent the two branches of STFP. $\mathrm{DW}_3(\cdot)$ denotes a $3 \times 3 \times 3$ depthwise separable convolution. $\mathrm{LN}(\cdot)$ denotes layer normalization. $\mathrm{ReLU}(\cdot)$ is an activation function. $\mathrm{PW}(\cdot)$ denotes pointwise convolution.

3.3. Efficient Bidirectional Sharing (EBS)

For the input sequence X, Efficient Bidirectional Sharing (EBS) uses different linear layers to separate forward features and backward features, which are then processed by SSMs modules with shared parameters. Subsequently, these two features are processed by two independent gating layers respectively, and finally concatenated into the output sequence. EBS suppresses specific dynamic features by introducing directional bias. The network structure of EBS is shown in Figure 4.

Specifically, EBS first normalizes the input sequence X, then linearly projects it into F_x, F_z, B_x, and B_z of the same dimension. Among them, F_x and F_z are used for the forward propagation branch, and B_x and B_z are used for the backward propagation branch. Then, pointwise convolution operations with shared parameters are applied to F_x and B_x of the forward and backward branches respectively, and then the output features are linearly projected for SSMs calculation with shared parameters. Finally, they are gated and combined by the corresponding F_z and B_z after activation functions, and the output sequence is obtained by linear projection to restore the dimension. For the SSMs in EBS, we adopt the

same default hyperparameters as Mamba. Among them, the state dimension is set to 16, and the expansion ratio is set to 2. EBS can be expressed as

$$F_{\text{forward}} = \text{SSM}^{\text{s}}\left(\text{PW}^{\text{s}}\left(\text{Lin}\left(\text{LN}(X)\right)\right)\right) \otimes \text{SILU}\left(\text{Lin}\left(\text{LN}(X)\right)\right), \tag{5}$$

$$B_{\text{backward}} = \text{SSM}^{\text{s}}\left(\text{PW}^{\text{s}}\left(\text{Lin}\left(\text{LN}(X)\right)\right)\right) \otimes \text{SILU}\left(\text{Lin}\left(\text{LN}(X)\right)\right), \tag{6}$$

$$\text{EBS} = \text{Lin}\left(F_{\text{forward}} + B_{\text{backward}}\right), \tag{7}$$

where $\text{LN}(\cdot)$ denotes layer normalization. $\text{Lin}(\cdot)$ is a linear layer. PW^{s} and SSM^{s} denote pointwise convolution and state space model with shared parameter weights respectively. $\text{SILU}(\cdot)$ is an activation function. $\otimes$ denotes element-wise multiplication.

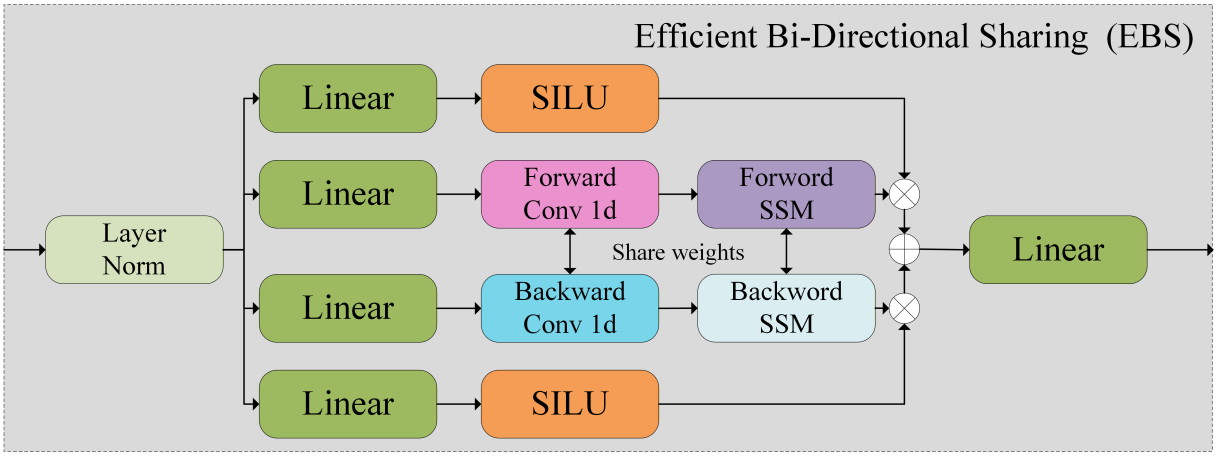


Figure 4. The Efficient Bidirectional Sharing (EBS) module.

3.4. Spatiotemporal Collaborative Modulation (STCM)

In our architecture design, we aim to expand the spatial information captured by ETMamba, enabling it to integrate both global and local modeling perspectives. This innovative idea gave birth to the new STCM Mamba Block module. Among them, the lower branch adopts Efficient Bidirectional Sharing (EBS) specifically responsible for extracting global temporal modeling, which achieves this goal by analyzing and calculating the long-range dependencies of feature maps. The upper branch focuses on capturing local spatial information at multiple scales. The local spatial modeling mechanism extracts contextual relationships by analyzing spatial patterns of different scales. The network structure of Spatiotemporal Collaborative Modulation (STCM) is shown in Figure 5.

Specifically, for the input sequence X, a normalization layer is first used to standardize the data. After standardization, the data passes through a local spatial modeling layer, which performs dimensionality reduction processing through a 1×1 convolution kernel. By quadrupling the number of channels, a more efficient tensor is generated for subsequent processing. To capture spatial context, we adopt multi-scale local spatial modeling that uses three dilated convolutions (DConv) with dilation rates of 1, 2, and 3, and convolution kernel sizes of 3×3, 5×5, and 7×7 respectively, enabling the model to capture both short-range and long-range spatial dependencies within each individual frame. After convolution processing, the feature map is activated by the SILU nonlinear function. Subsequently, the output features are residually connected with the original input to generate a tensor containing rich spatial information with specific dimensions. Finally, the outputs of the local spatial module branch and the global temporal module branch are concatenated along the channel dimension. This concatenation method integrates local and global spatial

information, enabling the model to fully understand the input feature maps. STCM can be expressed as

$$X' = \mathrm{PW}\big(\mathrm{LN}(X)\big), \tag{8}$$

$$X_C = \mathrm{concat}\big[\mathrm{DC}_3(X'), \mathrm{DC}_5(X'), \mathrm{DC}_7(X')\big] + X', \tag{9}$$

$$\mathrm{STCM} = \mathrm{concat}\big[\mathrm{EBS}, \big(\mathrm{LN}\big(\mathrm{SILU}(X_C)\big) + X\big)\big], \tag{10}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization. $\mathrm{PW}(\cdot)$ denotes pointwise convolution. $\mathrm{DC}_m(\cdot)$ denotes a dilated convolution with a kernel size of m. $\mathrm{SILU}(\cdot)$ is an activation function. $\mathrm{concat}[\cdot,\cdot]$ denotes concatenation along the tensor channel. EBS is Efficient Bidirectional Sharing.

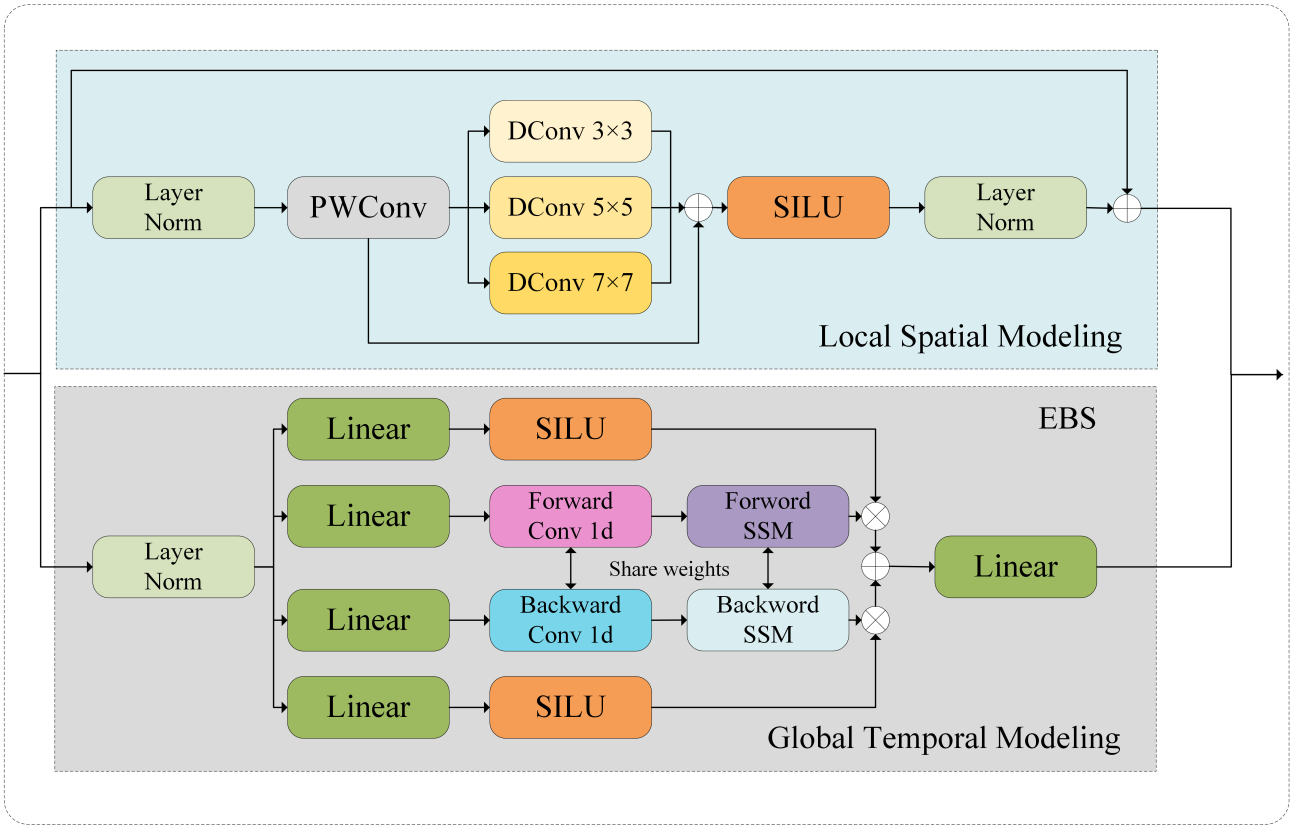


Figure 5. The Spatiotemporal Collaborative Modulation (STCM) module.

4. Results

4.1. Dataset

In this paper, we comprehensively evaluate the performance of our method on four benchmark video action recognition datasets, which collectively represent various scenarios and technical challenges in practical applications. These datasets are carefully selected to verify the robustness and adaptability of our proposed method under different video features:

(a) Kinetics-400 (K400) [4]: This dataset captures human activity videos from moving camera perspectives, including 240 K videos for training and 20 K videos for validation. The dataset contains 400 different categories, and each video has a duration of approximately 10 s. The diversity of action categories and motion scenarios makes it a challenging temporal modeling benchmark.

(b) Something-Something V2 (SSv2) [5]: This dataset is recorded with a fixed camera, containing 169 K training videos and 25 K validation videos, focusing on interactions between humans and objects. The dataset includes 174 action categories with an aver-

age video duration of 4 s. A large number of first-person perspectives and multi-object interactions in the dataset add additional complexity.

(c) HMDB-51 [6]: This dataset contains 6849 videos covering 51 action categories, with an average duration of video clips of approximately 5 s. It focuses on examining the model's ability to recognize temporal relationships and fine-grained differences between action categories from scarce data and limited training samples in a constrained environment.

(d) Breakfast [7]: This dataset is collected from real scenes of 18 different kitchens, containing 1712 video clips, and the action categories cover 10 typical breakfast-making operations. Different from other short video datasets, the average video duration of the Breakfast dataset is approximately 2.7 min. Moreover, each video clip in the Breakfast dataset is annotated at the action unit level, forming a complete hierarchical annotation that accurately captures the temporal dependencies and logical correlations of action units in daily human activities.

4.2. Model Configurations

All data processing, model construction, optimization, and training of the proposed ETMamba model in this paper are completed based on the PyTorch v2.1 and CUDA v11.8. framework, and the hardware environment for experimental training is an NVIDIA RTX3090 GPU. To balance computational efficiency and feature representation capability, the model selects VideoMamba-M as the backbone architecture. For different experimental datasets, the pre-trained weights of the ImageNet-1K [27] or K400 dataset are first used for initialization, then fine-tuned on the training set of each target dataset, and the final experimental results are reported on the validation set. In terms of data processing, during training and inference, we uniformly crop the video frame resolution to 224×224 through center cropping, then uniformly sample 16 frames with a time step of 2 for model input. The data augmentation strategy combines spatial, temporal, and advanced technologies, including label smoothing (coefficient 0.1), RandAug (parameters [9, 0.5]), random erasure, etc. Among them, flip augmentation is not enabled on the SSv2 dataset. In the inference phase, we adopt a multi-view cropping strategy for video data, and the final prediction score is obtained by calculating the average of the prediction scores of each view. The overall training scheme refers to the original VideoMamba architecture, and BFloat16 precision is used to improve training stability. Our model uses the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 0.05) for optimization training. The learning rate scheduling adopts a cosine decay learning rate scheduler, the initial learning rate is set to 3×10^{-4}, and the linear scaling strategy of the learning rate is consistent with VideoMamba. The batch size is set to 64, and Drop Path is set to 0.1. Regarding the number of learning rate warm-up epochs, K400 and SSV2 are 1 epoch, HMDB-51 and Breakfast are 5 epochs; the training epochs for each dataset are 30 epochs for K400, 35 epochs for SSV2, and 50 epochs for both HMDB-51 and Breakfast. We evaluate the ETMamba model on multiple video benchmark datasets and compare it with other advanced action recognition methods including video CNNs, Transformer, and Mamba.

4.3. Contrast Experiment

To verify the performance superiority and competitiveness of our proposed method in short video human action recognition tasks, we select the widely used benchmark dataset Kinetics-400 in the field of action recognition to conduct systematic comparative experiments. The comparative experimental results of ETMamba on the Kinetics-400 dataset are shown in Table 1.

Table 1. Comparison with the current state-of-the-art method on the K400 dataset.

Method	Pre-Training	Params	Fr. × Cr. × Cl.	FLOPs	Top-1	Top-5
ActionCLIP-B/16 [28]	CLIP-400M	142M	32 × 3 × 10	16.89T	83.8%	96.2%
UniFormerV2-B/16 [19]	CLIP-400M	115M	8 × 3 × 4	1.8T	85.6%	97.0%
AIM-L/14 [29]	CLIP-400M	341M	8 × 3 × 1	0.93T	86.8%	97.2%
DUALPATH-B/16 [30]	CLIP-400M	96M	32 × 3 × 1	2.1T	85.4%	97.1%
MoTE-L/14 [31]	CLIP-400M	346.6M	8 × 3 × 4	7.788T	86.8%	97.5%
MoTED-B/16 [32]	CLIP-400M	116M	32 × 3 × 1	2.04T	86.2%	97.5%
VideoMamba-M [3]	IN-1K	74M	16 × 3 × 4	2.424T	81.9%	95.4%
StableMamba-M [21]	IN-1K	76M	16 × 3 × 4	2.472T	82.5%	-
VideoMambaPro-Ti [20]	IN-1K	7M	32 × 3 × 4	0.4T	85.8%	96.9%
ViViM-S [33]	IN-1K	26M	16 × 3 × 4	0.816T	80.1%	94.1%
VideoMamba++ [34]	IN-1K	21M	64 × 3 × 4	1.62T	83.3%	96.0%
VideoMAP-M [22]	IN-1K	96M	16 × 3 × 4	-	85.8%	97.3%
VCMamba [35]	-	79M	- × 3 × 4	0.261T	87.3%	97.8%
MoMa-L/14 [36]	CLIP-400M	342M	16 × 1 × 3	4.152T	87.8%	98.0%
ETMamba(Ours)	IN-1K	97M	16 × 3 × 4	1.704T	**88.3%**	**98.5%**

In the table, 'Fr.', 'Cr.', and 'Cl.' are abbreviations for 'Frames', 'spatial crops', and 'temporal clips', respectively. The dash '-' indicates that the result is not provided in the reference literature. ETMamba achieves the optimal performance among all comparative models with a Top-1 accuracy of 88.3% and a Top-5 accuracy of 98.5%. Under the same configuration, compared with the baseline model VideoMamba (81.9% Top-1) which also relies on IN-1K pre-training, our method improves the Top-1 accuracy by 6.4 percentage points with fewer FLOPs (1.704T vs. 2.424T). More importantly, with a medium scale of only 97M parameters, ETMamba not only surpasses DUALPATH (96M, 85.4% Top-1) of the same parameter level by 2.9 percentage points but also outperforms the large-parameter model MoTE by 1.5 percentage points in accuracy, becoming the top-performing model among all comparative models. This performance not only demonstrates its efficient structural design but also indicates that when scaled to larger parameter levels in the future, the performance still has room for continuous improvement, showing excellent scalability. In terms of computational complexity, ETMamba's FLOPs are only 1.704T, which is at a low to medium level among comparative models. On the premise of ensuring the integrity of spatiotemporal information, it successfully reduces redundant computations and achieves the optimal trade-off between low overhead and high performance. In addition, although the accuracy of our method is similar to that of the model MoMa (342M, 87.8% Top-1), it is worth noting that it has far more adjustable parameters (97M vs. 342M) and larger FLOPs (1.704T vs. 4.152T) than our model. By virtue of its exquisite spatiotemporal modeling mechanism, ETMamba achieves performance surpassing large-parameter models at a medium parameter level, fully confirming the advancement and rationality of its structural design. At the same time, its low computational overhead makes it more suitable for real-time recognition practical application scenarios, with extremely high landing value. To more intuitively highlight the performance advantages of ETMamba, we draw a correlation diagram between the Top-1 accuracy and floating-point operations per video (FLOPs/Video) of various mainstream models on the Kinetics-400 dataset. Among them, our proposed method ETMamba is highlighted with a red five-pointed star, and the specific results are shown in Figure 6.

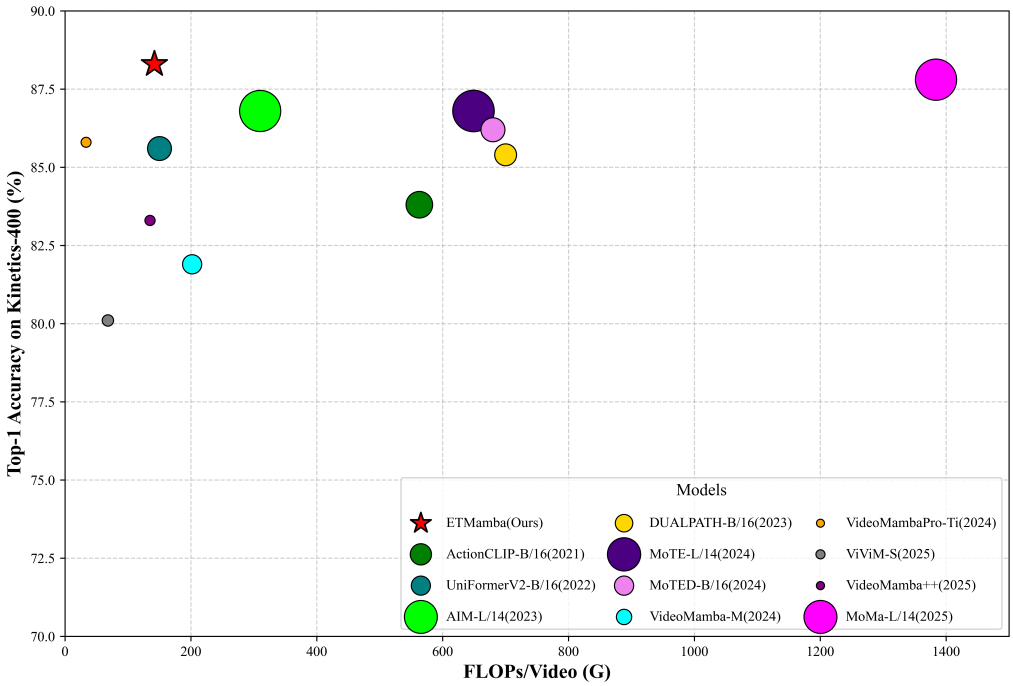


Figure 6. The performance of different models on the K400 dataset.

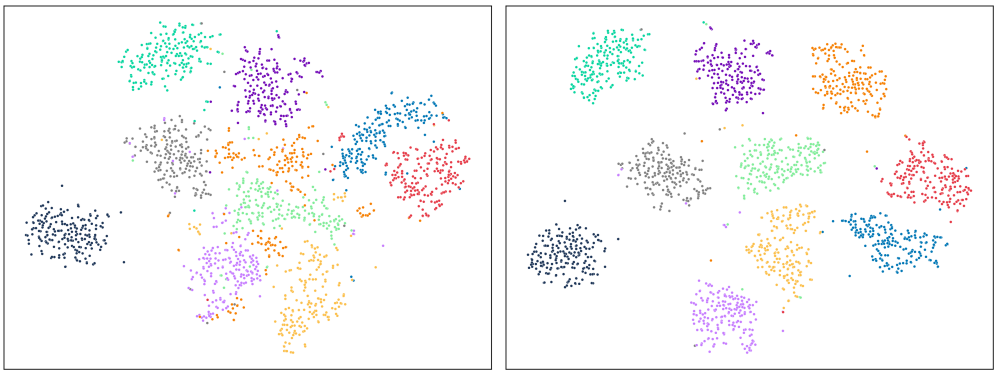
We also conducted experiments on the Something-Something V2 dataset and compared it with other advanced models in the field of action recognition in recent years. The comparative experimental results on the Something-Something V2 dataset are shown in Table 2.

Table 2. Comparison with the current state-of-the-art method on the SSv2 dataset.

Method	Pre-Training	Params	Fr. × Cr. × Cl.	FLOPs	Top-1	Top-5
UniFormerV2-B/16 [19]	CLIP-400M	163M	16 × 3 × 1	0.6T	69.5%	92.3%
UMT-B_{800e} [37]	CLIP-400M	87M	8 × 3 × 2	1.08T	70.8%	92.6%
PosMLP-Video-L [38]	-	35.4M	16 × 3 × 1	0.338T	70.3%	92.3%
MoTED-B [32]	CLIP-400M	112M	32 × 3 × 1	2.04T	71.9%	92.7%
VideoMamba-M [3]	IN-1K	74M	16 × 3 × 4	2.424T	68.3%	91.4%
VideoMamba++ [34]	IN-1K	16M	56 × 3 × 4	0.672T	69.6%	92.2%
MoMa-L/14 [36]	CLIP-400M	342M	16 × 1 × 3	8.304T	73.8%	93.6%
ETMamba(Ours)	IN-1K	97M	16 × 3 × 4	1.704T	**74.6%**	**94.1%**

In the table, 'Fr.', 'Cr.', and 'Cl.' are abbreviations for 'Frames', 'spatial crops', and 'temporal clips', respectively. The dash '-' indicates that the result is not provided in the reference literature. Our proposed ETMamba method comprehensively outperforms several recently proposed methods. Compared with MoTED and VideoMamba in 2024, it achieves a leap in Top-1 accuracy of 2.7% and 6.3% respectively. Even compared with MoMa proposed in 2025, our method achieves a 0.8% improvement in Top-1 accuracy. Although the improvement is not high, its parameter scale (342M vs. 97M) is 3.5 times that of ETMamba, and the computational volume (8.304T vs. 0.853T) is nearly 10 times higher, which limits its practicality in actual deployment scenarios. Although VideoMamba++ has better computational efficiency than our model, it relies on more video frame inputs (56 Frames vs. 16 Frames), and its Top-1 accuracy is much lower than our method (69.6% vs. 74.6%). This indicates that ETMamba can efficiently capture key temporal information of actions within a more compact input temporal length, achieving a balance between performance and efficiency. Furthermore, as shown in Figure 7, we randomly selected 10 action categories from the SSv2 dataset and performed visual analysis of the model-

extracted features using the T-Distributed Stochastic Neighbor Embedding (t-SNE) [39] method. The comparison results clearly reveal the significant difference in feature learning ability between the baseline method and our proposed ETMamba method.The feature distribution generated by the baseline model is relatively loose, the feature difference of samples in the same class is large, and the boundary between different classes is not clear. In stark contrast, the features extracted by the ETMamba method demonstrate a more compact clustering structure, with distinct clusters of different categories and higher separation between them. This outstanding clustering performance demonstrates that ETMamba effectively decouples the complex spatiotemporal information embedded in video sequences, enabling a clear differentiation of various action categories in the feature space.



(**a**) (**b**)

Figure 7. The t-SNE visualization of 10 randomly selected categories in the SSv2 dataset, where different colors represent distinct behavioral categories, with each point corresponding to a feature embedding of a video sequence. (**a**) Vanilla model. (**b**) ETMamba(Ours).

Beyond performance comparison, we further analyze the computational complexity of our ETMamba in comparison with the baseline VideoMamba. We provide a detailed analysis of the computational complexity of ETMamba. Built upon VideoMamba as the baseline, ETMamba has a slight increase in the number of parameters from 74 million to 97 million, which is mainly caused by three newly introduced lightweight modules: STFP, EBS, and STCM. All these modules are elaborately designed with low computational overhead and no high-cost operations. Specifically, STFP adopts depthwise separable convolution with linear complexity; EBS achieves bidirectional temporal modeling through parameter-sharing state space models; and STCM captures multi-scale spatial features via sparse dilated convolutions. Despite the minor growth in parameters, ETMamba only costs 1.704 TFLOPs, which is 29.6% lower than the baseline VideoMamba (2.424 TFLOPs). Such a remarkable improvement in computational efficiency benefits from the targeted spatiotemporal optimization of each module, which effectively reduces invalid and redundant feature computations during the spatiotemporal modeling process. Therefore, ETMamba achieves a favorable trade-off between modeling performance and computational cost.

In addition, we also conducted experiments on the small-scale action recognition dataset HMDB-51 to study the generalization ability of the spatiotemporal and action features learned by our method. The comparative experimental results on the HMDB-51 dataset are shown in Table 3.

Table 3. Comparison with the current state-of-the-art method on the HMDB-51 dataset.

Method	Pre-Training	Params	Top-1
VideoMAE [40]	K400	87M	73.3%
VideoMAE V2 [41]	-	1050M	**88.1%**
VideoMamba-M [3]	K400	74M	68.6%
VideoMambaPro-M [20]	IN-1K	72M	63.2%
Mamba-ND [24]	IN-1K	36M	60.9%
ETMamba(Ours)	K400	97M	75.7%

In the table, the dash '-' indicates that the result is not provided in the reference literature. Through comparison, it can be found that ETMamba shows excellent competitiveness among many models. Although its Top-1 accuracy is lower than that of VideoMAE V2 (75.7% vs. 88.1%), its parameter quantity is an order of magnitude higher than that of ETMamba. Additionally, we present the confusion matrix for the HMDB-51 dataset, as shown in Figure 8. The visualization results show that our model performs well in most action categories. ETMamba achieves an accuracy of no less than 70% in all 51 categories. This widespread high-accuracy performance far exceeds the limitation of most models that excel in some categories but are mediocre in most categories. Moreover, the model's temporal feature capture is good. The recognition accuracy of limb posture recognition that requires previous and subsequent frames, such as stand (79.31%) and sit (80.56%), run (79.07%) and walk (78.57%), is all close to 80%. For facial expression categories that require local spatial fine-grained recognition, such as smile (81.25%) and laugh (78.38%), the model can also accurately distinguish them, fully demonstrating ETMamba's collaborative capture capability in global time and local space.

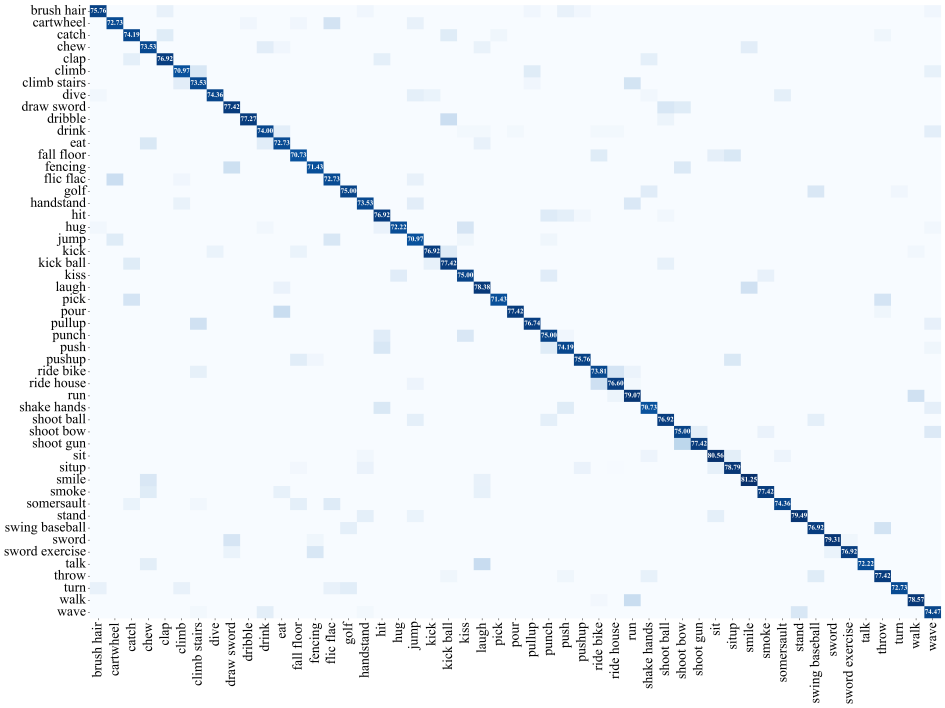


Figure 8. HMDB-51 confusion matrix. The confusion matrix displays predicted labels on the *x*-axis and actual labels on the *y*-axis, with the diagonal numbers representing the corresponding Top-1 accuracy. The intensity of the color in the matrix corresponds to the accuracy of the prediction.

To further verify the effectiveness of this method in capturing long video sequences, we also conducted experiments on the Breakfast dataset and compared it with other

advanced methods in the field of action recognition. The comparative experimental results of ETMamba on the Breakfast dataset are shown in Table 4.

Table 4. Comparison with the current state-of-the-art method on the Breakfast dataset.

Method	End-to-End	Pre-Training	Frames	Top-1
Distant Supervision [42]	×	IN-21K+HTM	-	89.9%
ViS4mer [43]	×	IN-21K+K600	-	88.2%
Turbo [44]	×	K400+HTM-AA	32	91.3%
VideoMamba-M [3]	✓	K400	64	95.8%
MoMa-L/14 [36]	✓	K400+CLIP-400M	64	96.9%
VideoMAP-M [22]	✓	K400	24	97.9%
ETMamba(Ours)	✓	K400	64	**98.1%**

In the table, 'End-to-End' indicates an end-to-end method, where '✓' and '×' indicate whether it is adopted. The dash '-' indicates that the result is not provided in the reference literature. Our ETMamba model performs excellently on the Breakfast dataset, achieving a Top-1 accuracy of 98.1%, which is a 2.3% improvement compared with the baseline model VideoMamba. This method adopts an end-to-end learning paradigm, which adaptively learns spatiotemporal features and behavioral patterns directly from original video frames, ensuring both the integrity of long-sequence temporal information and the consistency and effectiveness of feature mapping. This architectural advantage enables it to overwhelmingly outperform non-end-to-end methods such as Turbo (98.1% vs. 91.3%), strongly confirming the core value of the end-to-end paradigm in long video understanding tasks. It is worth noting that compared with the MoMa model that relies on multi-source pre-training data, ETMamba can deeply mine fine-grained spatiotemporal correlation information in videos relying only on the single pre-training dataset K400, and still achieve stronger generalization ability without relying on massive multi-source data, fully demonstrating its effectiveness in the field of long video understanding.

4.4. Ablation Experiments

To verify the effectiveness of each core module of the ETMamba model, we systematically explored the effects of the Spatiotemporal Feature Preservation (STFP), Efficient Bidirectional Sharing (EBS), and Spatiotemporal Collaborative Modulation (STCM) modules. The experiments were conducted on the K400 and SSv2 datasets, with performance evaluated on their respective validation sets. The parameters were consistent with the main experiments, and 16 frames were sampled for each video clip. The ablation experiment results are presented in Table 5.

Table 5. Comparison of performance of each module in ETMamba.

	STFP	EBS	STCM	Top-1 (K400)	Top-1 (SSv2)
(a)	×	×	×	81.9%	68.3%
(b)	✓	×	×	85.1% (+3.2)	70.7% (+2.4)
(c)	✓	✓	×	86.4% (+1.3)	72.8% (+2.1)
(d)	✓	✓	✓	88.3% (+1.9)	74.6% (+1.8)

(a) Baseline model: We use VideoMamba as the baseline model for ablation experiments, which has a Top-1 accuracy of 81.9% on the Kinetics-400 dataset and 68.3% on the Something-Something V2 (SSv2) dataset, providing a benchmark reference for verifying the effectiveness of subsequent modules.

(b) Introducing only the STFP module: We add the Spatiotemporal Feature Preservation (STFP) module to the baseline model. This module combines depthwise separable

convolution and a gating fusion mechanism to extract and preserve spatiotemporal features from video data before sequence flattening, which can effectively retain the foreground spatial features of the appearance-biased Kinetics-400 (K400) dataset and the motion temporal features of the temporal-sensitive Something-Something V2 (SSv2) dataset, thus maintaining the complete spatial position correlation and temporal inter-frame information of the original video data. The experimental results show that after introducing STFP, the model's Top-1 accuracy increases from 81.9% to 85.1% on K400 (an increase of 3.2%) and from 68.3% to 70.7% on SSv2 (an increase of 2.4%). This result fully proves that the STFP module solves the problem of insufficient feature integrity caused by the flattening operation of the traditional Mamba model, providing more accurate feature input for subsequent spatiotemporal modeling.

(c) Introducing STFP + EBS modules: On the basis of the STFP module, we further added the Efficient Bidirectional Sharing (EBS) strategy as a new type of SSMs block. EBS achieves accurate modeling of bidirectional temporal dependencies through a directional bias mechanism and parameter sharing design. The experimental results show that after adding EBS, the model's Top-1 accuracy is further increased by 1.3% on K400 (85.1% vs. 86.4%) and by 2.1% on SSv2 (70.7% vs. 72.8%), indicating that the EBS module can effectively enhance bidirectional key information, weaken invalid dynamics, and enable the model to more accurately anchor the temporal correlation of input features. The more significant performance gain on SSv2 aligns with our design intuition: the semantics of most actions in SSv2 are determined by temporal direction, and the EBS bidirectional temporal modeling capability is particularly critical for this dataset that focuses on inter-frame temporal motion relationship modeling.

(d) Introducing STFP + STCM modules: We further replace the standalone EBS block with the complete Spatiotemporal Collaborative Modulation (STCM) mechanism that inherently integrates the EBS structure within its dual-branch framework. STCM uses a dual-branch division of labor architecture of "global temporal–local spatial", combined with multi-scale dilated convolution to enhance fine-grained spatial feature extraction capability. Due to the serpentine row-major scanning strategy of Mamba, this design specifically provides additional spatiotemporal information compensation for feature extraction in 1D sequences, achieving collaborative optimization of global temporal flow capture and local spatial detail mining. The experimental results show that the model's Top-1 accuracy is increased from 86.4% to 88.3% on K400 (a relative increase of 1.9%) and from 72.8% to 74.6% on SSv2 (a relative increase of 1.8%), fully proving that the STCM module can effectively collaboratively model spatiotemporal information, significantly enhance the model's ability to capture fine-grained behavioral features, and further improve recognition accuracy.

The results of the ablation experiments show that by sequentially introducing the three core modules of STFP, EBS, and STCM, the Top-1 accuracy of ETMamba is gradually increased from 81.9% of the baseline model to 88.3% on K400, and from 68.3% to 74.6% on SSv2. The addition of each module significantly improved the performance of the model, and each step of improvement specifically solved the problems of the Mamba model such as video data feature loss, vague capture of directional features, and insufficient capture of local fine-grained features. This result effectively proves the effectiveness of the proposed model, making ETMamba a high-performance model that achieves an excellent balance between efficiency and effectiveness.

5. Conclusions

To address the balance problem between spatiotemporal modeling accuracy and computational efficiency in existing video action recognition methods, this paper proposes an efficient temporal model—ETMamba. By performing targeted optimization of the

traditional Mamba architecture, the model systematically solves the core limitations of Mamba in video tasks. The main research results are as follows.

Firstly, to address the problems of broken spatial relationships and loss of temporal information caused by data flattening in the Mamba model, the Spatiotemporal Feature Preservation (STFP) module is designed. This module completes the extraction and fusion of original spatiotemporal features before block division through depthwise separable convolution and a gating fusion mechanism, providing more accurate feature input for subsequent modeling. Secondly, to break through the limitation of unidirectional processing mode on the capture of directional features, the Efficient Bidirectional Sharing (EBS) strategy is proposed. By splitting forward and backward branches and sharing the core parameters of SSMs, a directional bias mechanism is introduced to enhance key temporal information, achieving accurate modeling of bidirectional temporal dependencies. Finally, the Spatiotemporal Collaborative Modulation (STCM) mechanism is constructed, which adopts a dual-branch architecture of global temporal and local spatial collaboration, combined with multi-scale dilated convolution to enhance fine-grained spatial feature extraction capability, effectively balancing the needs of long-short term dependency modeling and significantly improving the model's ability to distinguish behaviors in complex scenarios. By combining the above designs, our model efficiently completes the extraction and modeling of video spatiotemporal information with low computational cost, significantly enhancing the model's interpretability and generalization ability. Thus, excellent model performance is achieved on various real-world datasets. Finally, we further verified the excellent recognition performance and generalization ability of ETMamba on long and short video datasets through comparative experiments, ablation experiments, etc. On large-scale datasets Kinetics-400 and Something-Something V2, its accuracy surpasses models of the same parameter level and some large-parameter models, and the computational complexity is at a low to medium level. The model can also adapt to small-scale datasets and long video datasets, which is verified by our experiments on HMDB-51 and Breakfast datasets. Although ETMamba has achieved significant performance improvements, it still has certain limitations. Although the model maintains low to medium computational complexity, it still has room for compression compared with lightweight models, and the inference overhead needs to be further reduced in resource-constrained real-time scenarios. Currently, the model is mainly designed for action classification tasks and has not been extended to more complex video analysis tasks such as action detection and temporal localization. In addition, we will try to pre-train on larger-scale datasets and adopt larger benchmark models to further explore the model's representation capability and promote its application in a wider range of practical scenarios.

Author Contributions: Conceptualization, C.W.; methodology, R.H.; software, R.H.; validation, R.H., C.L., L.Z., Z.N., Y.S., H.S., M.L., H.C. and P.S.; formal analysis, C.W.; investigation, R.H., C.L., L.Z., Z.N. and Y.S.; resources, H.S., H.C., P.S. and M.L.; data curation, C.W.; writing—original draft preparation, R.H.; writing—review and editing, C.W.; visualization, R.H.; supervision, C.W., H.S., H.C., P.S. and M.L.; project administration, C.W.; funding acquisition, C.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the research and planning project of Jilin Provincial Department of Education (No.JJKH20240441HT, JJKH20220376SK), Key Laboratory open fund of Modern Agricultural Equipment and Technology (Jiangsu University), Ministry of Education (No.MAET202315), The Industrial Technology and Development Project of Development and Reform Commission of Jilin Province (No.2021C044-8, 2023C030-3), Key Laboratory open fund of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources (No.KF-2021-06-067), Jilin Provincial Science and Technology Development Plan Project (No.20210203013SF), and the National Natural Science Foundation of China (Key Program) (No.62376106).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: Thanks to all who contributed to this paper.

Conflicts of Interest: Author Mingqi Li was employed by Midea Group (Shanghai) Co., Ltd., Qingpu District, Shanghai, China. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LO, USA, 19–20 June 2022; pp. 11963–11975.
2. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417. [CrossRef]
3. Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; Qiao, Y. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2024; pp. 237–255.
4. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.06950. [CrossRef]
5. Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: Piscatway, NJ, USA, 2017; pp. 5842–5850.
6. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*; IEEE: Piscataway, NJ. USA, 2011; pp. 2556–2563.
7. Kuehne, H.; Arslan, A.; Serre, T. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscatway, NJ, USA, 2014; pp. 780–787.
8. Lin, J.; Gan, C.; Han, S. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscatway, NJ, USA, 2019; pp. 7083–7093.
9. Zhang, Y.; Bai, Y.; Wang, H.; Xu, Y.; Fu, Y. Look more but care less in video recognition. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 30813–30825.
10. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscatway, NJ, USA, 2019; pp. 6202–6211.
11. Wang, H.; Xia, T.; Li, H.; Gu, X.; Lv, W.; Wang, Y. A Channel-Wise Spatial-Temporal Aggregation Network for Action Recognition. *Mathematics* **2021**, *9*, 3226. [CrossRef]
12. Xia, L.; Fu, W. Spatial-temporal multiscale feature optimization based two-stream convolutional neural network for action recognition. *Clust. Comput.* **2024**, *27*, 11611–11626. [CrossRef]
13. Bertasius, G.; Wang, H.; Torresani, L. Is space-time attention all you need for video understanding? *Int. Conf. Mach.* **2021**, *2*, 4.
14. Zhang, Y.; Li, X.; Liu, C.; Shuai, B.; Zhu, Y.; Brattoli, B.; Chen, H.; Marsic, I.; Tighe, J. Vidtr: Video transformer without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 13577–13587.
15. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscatway, NJ, USA, 2022; pp. 3202–3211.
16. Li, Y.; Wu, C.Y.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; Feichtenhofer, C. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscatway, NJ, USA, 2022; pp. 4804–4814.
17. Li, S.; Wang, Z.; Liu, Y.; Zhang, Y.; Zhu, J.; Cui, X.; Liu, J. FSformer: Fast-Slow Transformer for video action recognition. *Image Vis. Comput.* **2023**, *137*, 104740. [CrossRef]
18. Li, K.; Wang, Y.; Zhang, J.; Gao, P.; Song, G.; Liu, Y.; Li, H.; Qiao, Y. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12581–12600. [CrossRef]
19. Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Wang, L.; Qiao, Y. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv* **2022**, arXiv:2211.09552.
20. Lu, H.; Salah, A.A.; Poppe, R. Videomambapro: A leap forward for mamba in video understanding. *arXiv* **2024**, arXiv:2406.19006.
21. Suleman, H.; Talal Wasim, S.; Naseer, M.; Gall, J. Distillation-free Scaling of Large SSMs for Images and Videos. *arXiv* **2024**, arXiv:2409.11867. [CrossRef]
22. Liu, Y.; Wu, P.; Liang, C.; Shen, J.; Wang, L.; Yi, L. VideoMAP: Toward Scalable Mamba-based Video Autoregressive Pretraining. *arXiv* **2025**, arXiv:2503.12332.

23. Beedu, A.; Dong, Z.; Sheinkopf, J.; Essa, I. Mamba Fusion: Learning Actions Through Questioning. In *ICASSP 2025—IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscatway, NJ, USA, 2025; pp. 1–5.
24. Li, S.; Singh, H.; Grover, A. Mamba-nd: Selective state space modeling for multi-dimensional data. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2024; pp. 75–92.
25. Gao, X.; Kanu-Asiegbu, A.M.; Du, X. Mambast: A plug-and-play cross-spectral spatial-temporal fuser for efficient pedestrian detection. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*; IEEE: Piscatway, NJ, USA, 2024; pp. 2027–2034.
26. Sheng, J.; Zhou, J.; Wang, J.; Ye, P.; Fan, J. DualMamba: A lightweight spectral–spatial mamba-convolution network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *63*, 5501415. [CrossRef]
27. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
28. Wang, M.; Xing, J.; Liu, Y. Actionclip: A new paradigm for video action recognition. *arXiv* **2021**, arXiv:2109.08472. [CrossRef]
29. Yang, T.; Zhu, Y.; Xie, Y.; Zhang, A.; Chen, C.; Li, M. Aim: Adapting image models for efficient video action recognition. *arXiv* **2023**, arXiv:2302.03024. [CrossRef]
30. Park, J.; Lee, J.; Sohn, K. Dual-path adaptation from image to video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscatway, NJ, USA, 2023; pp. 2203–2213.
31. Zhu, M.; Wang, Z.; Hu, M.; Dang, R.; Lin, X.; Zhou, X.; Liu, C.; Chen, Q. Mote: Reconciling generalization with specialization for visual-language to video knowledge transfer. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 55403–55424.
32. Zhang, W.; Wan, C.; Liu, T.; Tian, X.; Shen, X.; Ye, J. Enhanced motion-text alignment for image-to-video transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscatway, NJ, USA, 2024; pp. 18504–18515.
33. Yang, Y.; Xing, Z.; Yu, L.; Fu, H.; Huang, C.; Zhu, L. Vivim: A Video Vision Mamba for Ultrasound Video Segmentation. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 10293–10304. [CrossRef]
34. Song, X.; Tian, W.; Zhu, Q.; Zhang, X. VideoMamba++: Integrating state space model with dual attention for enhanced video understanding. *Image Vis. Comput.* **2025**, *161*, 105609. [CrossRef]
35. Hu, Y.; Zhao, J.; Qi, C.; Qiang, Y.; Zhao, J.; Pei, B. VC-Mamba: Causal Mamba representation consistency for video implicit understanding. *Knowl.-Based Syst.* **2025**, *317*, 113437. [CrossRef]
36. Yang, Y.; Ma, C.; Mao, Z.; Yao, J.; Zhang, Y.; Wang, Y. MoMa: Modulating Mamba for Adapting Image Foundation Models to Video Recognition. *arXiv* **2025**, arXiv:2506.23283. [CrossRef]
37. Li, K.; Wang, Y.; Li, Y.; Wang, Y.; He, Y.; Wang, L.; Qiao, Y. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscatway, NJ, USA, 2023; pp. 19948–19960.
38. Hao, Y.; Zhou, D.; Wang, Z.; Ngo, C.W.; Wang, M. Posmlp-video: Spatial and temporal relative position encoding for efficient video recognition. *Int. J. Comput. Vis.* **2024**, *132*, 5820–5840. [CrossRef]
39. Maaten, L.V.D.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
40. Tong, Z.; Song, Y.; Wang, J.; Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 10078–10093.
41. Wang, L.; Huang, B.; Zhao, Z.; Tong, Z.; He, Y.; Wang, Y.; Qiao, Y. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition;* IEEE: Piscataway, NJ, USA, 2023; pp. 14549–14560.
42. Lin, X.; Petroni, F.; Bertasius, G.; Rohrbach, M.; Chang, S.F.; Torresani, L. Learning to recognize procedural activities with distant supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscatway, NJ, USA, 2022; pp. 13853–13863.
43. Islam, M.M.; Bertasius, G. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2022; pp. 87–104.
44. Han, T.; Xie, W.; Zisserman, A. Turbo training with token dropout. *arXiv* **2022**, arXiv:2210.04889. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.