electronics

MDPI

Article

RoPE-T: Rotary Position Embedding for Enhanced Temporal Modeling in Video-Based Human Action Recognition

Fatemeh Shirazi [1,2] and Mohammad Taghi Sadeghi [1,*]

[1] Department of Electrical Engineering, Yazd University, Yazd P.O. Box 8915818411, Iran
[2] Department of Computer Science, University of Guelph, Guelph, ON N1G 2W1, Canada
* Correspondence: m.sadeghi@yazd.ac.ir

Abstract

Human action recognition requires modeling both spatial information and temporal relationships across video frames. While Transformer encoders are effective for capturing long-range dependencies, standard self-attention mechanisms do not inherently encode the sequential order of input tokens. To address this limitation, we introduce RoPE-T, a RoPE-enhanced temporal Transformer framework for video-based human action recognition. Our approach employs a pretrained EfficientNetV2B0 as a lightweight spatial feature extractor, followed by a temporal Transformer encoder. We apply Rotary Position Embedding (RoPE) directly to the query and key projections within the self-attention module, allowing relative temporal position information to influence the attention scores without modifying the spatial backbone. We also investigate the effect of the rotary base parameter on action recognition performance, showing that its impact varies across datasets and Transformer configurations. Evaluations on the UCF101 and HMDB51 datasets show improvements over our baseline model under identical architectural and training settings, increasing accuracy from 91.51% to 91.59% on UCF101 and from 74.51% to 75.10% on HMDB51. After tuning, the proposed RoPE-T model achieves accuracies of 91.91% on UCF101 and 78.82% on HMDB51, outperforming several previously published action recognition methods on both benchmarks. These results indicate that RoPE is a viable alternative for incorporating relative temporal positional information into CNN–Transformer models for human action recognition.

Keywords: human action recognition; video Transformer; Rotary Position Embedding (RoPE); temporal attention; CNN-Transformer

Check for updates

Academic Editor: Kefeng Ji

Received: 12 September 2026
Revised: 23 October 2026
Accepted: 23 October 2026
Published: 27 October 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

Human action recognition (HAR) is a core problem in computer vision and video understanding [1]. It is essential for many real-world applications, including intelligent surveillance [2], healthcare monitoring [3], sports analytics [4], human–computer interaction [5], video retrieval, and captioning [6]. However, HAR is difficult because action videos contain complex variations in viewpoint, background clutter, human pose, occlusion, and scene context [7–9]. While early deep learning approaches relied on two-stream [10] and 3D convolutional neural networks (CNNs) to learn spatiotemporal features [11,12], the introduction of the foundational Transformer architecture [13] has recently inspired a shift toward attention-based vision models capable of capturing long-range dependencies [14–16].

Prior to Transformers, models captured long-range temporal dynamics using segment-based architectures [17,18], sparse sampling, and recurrent neural networks (RNNs) such

as LSTMs [19,20]. These studies demonstrated that temporal continuity and the precise ordering of frames are just as important as spatial appearance. To better focus on these informative segments, researchers later integrated early attention mechanisms into convolutional and recurrent networks [21–24]. However, these sequential architectures were difficult to parallelize and required computationally expensive fusion.

Effective temporal modeling remains a major challenge. Human actions often depend on subtle motion transitions and the precise sequential ordering of frames [20,25]. Since standard self-attention treats tokens as an unordered set [26], video Transformers typically inject temporal order using learnable absolute positional embeddings. However, this additive approach only assigns a static identity to each token; it does not explicitly model the relative temporal distance between frames. While researchers have widely adopted hybrid CNN-Transformer architectures to balance accuracy and efficiency [27–29], many still rely on computationally heavy fusion strategies, optical flow streams, or complex 3D tokenization to compensate for poor temporal reasoning [30–32]. Consequently, the foundational design of positional encoding is often overlooked.

To address this limitation, we propose RoPE-T, a RoPE-enhanced temporal Transformer tailored for video-based human action recognition. Instead of relying on static absolute positional embeddings, we integrate Rotary Position Embedding (RoPE) [33] directly into the query and key projections of the multi-head self-attention module. This formulation allows relative temporal displacement to influence the attention scores without adding extra computational overhead. Following efficient hybrid architectures such as the Video Transformer Network (VTN) [34], we use a pretrained 2D CNN as an off-the-shelf spatial feature extractor and pass the resulting compact frame-level tokens to our temporal encoder.

While recent studies have explored RoPE for video understanding, they have primarily focused on large-scale Video–Language Models (VLMs) [35,36]. However, the use of RoPE within temporal self-attention for RGB-based human action recognition remains underexplored. This work focuses on a controlled empirical evaluation of RoPE within temporal self-attention for RGB-based human action recognition, comparing it with learnable absolute positional embeddings under matched architectural and training settings on UCF101 and HMDB51. We further investigate the rotary base parameter (θ) and show that the best-performing value among those evaluated varies depending on the dataset and model configuration.

The main contributions of this work are summarized as follows:

1. We propose RoPE-T, a RoPE-enhanced temporal Transformer for human action recognition. By applying Rotary Position Embedding directly to the query and key projections, the self-attention mechanism captures relative temporal distances without relying on static absolute embeddings.
2. We investigate the effect of the rotary base parameter (θ) on action recognition performance and show that its influence varies across datasets and Transformer configurations.
3. We evaluate the proposed framework on the UCF101 and HMDB51 datasets. Our experiments show modest accuracy improvements over the baseline under identical settings, while further hyperparameter tuning improves the final results. Temporal attention and Grad-CAM visualizations provide additional insights into the model's behavior across frames and spatial regions.

The rest of this paper is organized as follows. Section 2 reviews related works. Section refsec:method presents the details of our baseline architecture and the proposed RoPE integration. In Section 4, we report the experimental setup, quantitative results, ablation studies, and qualitative analyses. Finally, Section 5 concludes the paper.

2. Related Works

2.1. Video-Based HAR and Hybrid Architectures

Action recognition depends on both spatial semantics and temporal motion cues. Early deep learning models addressed this by processing RGB frames and optical flow through two-stream networks [10], by extracting spatiotemporal features directly via 3D convolutions [11,12]. While these methods provided strong baselines, they usually required dense video processing and high computational cost. To reduce redundant frame processing, temporal segment modeling demonstrated that representing a video through sparsely sampled segments can capture long-range structure [17,20]. Before attention mechanisms dominated, recurrent modules [19] were coupled with 2D CNNs for temporal aggregation, but their sequential nature severely limited parallelization.

More recently, hybrid CNN-Transformer architectures have emerged as a practical alternative [37,38]. Following the design paradigm of VTN [34], these lightweight architectures use 2D convolutional backbones to extract spatial features and a Transformer encoder to model temporal relationships across the resulting frame-level features [39–41]. In other words, these networks decouple spatial and temporal learning. Our framework adopts this decoupled architecture but focuses specifically on improving the temporal encoding mechanism without introducing additional modalities, optical flow, or computationally heavy 3D convolutions.

2.2. Attention Mechanisms and Video Transformers

Early attention mechanisms were designed to highlight discriminative temporal segments and suppress background noise [21,24]. Inspired by advancements in natural language processing, pure video Transformers, such as TimeSformer [42] and ViViT [14], tokenize videos across space and time to capture long-range dependencies [43]. Hierarchical architectures such as Video Swin Transformer [44] and MViTv2 [45], build upon this by extracting multi-scale spatiotemporal representations. Although fully spatiotemporal Transformers achieve high accuracy, their quadratic complexity in the number of tokens and reliance on extensive pretraining limit practical deployment. Recent studies emphasize the necessity of filtering out redundant spatial tokens and developing more efficient temporal aggregation modules [16,46].

Many methods attempt to improve temporal reasoning through multi-head attention [47], like two-stream collaboration, and motion-aware frame selection [48–51]. Our framework takes a different route: rather than adding architectural complexity through additional modules or streams, we enhance temporal modeling by refining the positional encoding mechanism directly within the self-attention dot product—a simpler and more fundamental intervention.

2.3. Positional Encoding in Transformers

Standard self-attention is permutation-invariant, requiring an explicit mechanism to encode sequence order [13]. Although the original Transformer used fixed sinusoidal encodings, modern vision models default to learnable absolute positional embeddings [26]. This additive strategy, however, only assigns a static identity to each token. It does not explicitly capture the relative temporal distances between frames, which are important for understanding complex human motion where the ordering and spacing of poses matter more than absolute frame indices. Relative positional methods, such as ALiBi [52] and decoupled relative embeddings [53], attempt to map pairwise distances directly into the attention matrix [54], yet they rely on heuristic additive bias terms rather than structural feature transformations.

Rotary Position Embedding (RoPE) [33] provides a different geometric approach, encoding position by rotating the query and key representations in paired dimensions. Despite its recent adoption in 2D spatial vision tasks [55] and large video–language models [35,36], the use of RoPE in lightweight, RGB-only video Transformers remains underexplored. Existing video-specific RoPE studies show that default frequency allocation can produce oscillatory effects that weaken long-range temporal reasoning, motivating our investigation into tuning the rotary base parameter for smoother temporal encoding. In this paper, we investigate this setting by integrating RoPE into the temporal self-attention module of a CNN–Transformer pipeline and analyzing the effect of its rotational frequency. Our experiments show that the effect of tuning θ on action recognition accuracy varies across datasets.

3. Proposed RoPE-Enhanced Video Transformer

3.1. Overall Architecture

The proposed framework, RoPE-T, is designed to model temporal relationships in video sequences. Rather than employing computationally heavy 3D tokenization, we use a hybrid CNN-Transformer architecture. The pipeline consists of three main stages: (1) spatial feature extraction using a pretrained 2D CNN, (2) temporal sequence modeling via a Transformer encoder enhanced with Rotary Position Embedding (RoPE), and (3) action classification. By integrating RoPE directly into the temporal self-attention mechanism, the model incorporates relative temporal displacement between video frames without relying on learnable absolute positional embeddings.

3.2. CNN-Based Spatial Feature Extraction

Given an input video, we uniformly sample a fixed-length sequence of T RGB frames, denoted as in Equation (1):

$$V = \{I_1, I_2, \ldots, I_T\} \tag{1}$$

We employ a transfer learning strategy using Efficient-NetV2 [56], pretrained on ImageNet, as an off-the-shelf feature extractor. This avoids updating millions of parameters while extracting discriminative spatial features. The backbone weights are kept frozen throughout training to provide a consistent spatial feature representation for both the baseline and RoPE-enhanced models. This design allows us to focus on the effect of temporal positional encoding without introducing additional variations from spatial feature fine-tuning.

For each sampled frame I_t, the CNN outputs a compact spatial representation:

$$f_t = \mathcal{F}_{\text{CNN}}(I_t) \tag{2}$$

where $\mathcal{F}_{\text{CNN}}(\cdot)$ represents the CNN backbone with its classification head removed, and $f_t \in \mathbb{R}^D$ is the extracted feature vector of dimension D. These frame-level descriptors are stacked sequentially to form a temporal sequence $X = [f_1, f_2, \ldots, f_T] \in \mathbb{R}^{T \times D}$. Before temporal modeling, a learnable linear projection is applied to map the feature sequence to the Transformer's hidden dimension d, ensuring compatibility with the subsequent attention layers:

$$Z = XW_e + b_e \tag{3}$$

where $W_e \in \mathbb{R}^{D \times d}$ and $b_e \in \mathbb{R}^d$ are learnable projection weights and biases, respectively. The resulting sequence $Z \in \mathbb{R}^{T \times d}$ serves as the input tokens for the temporal encoder.

3.3. Baseline Temporal Self-Attention

As illustrated in Figure 1, our baseline architecture follows a standard CNN-Transformer pipeline.

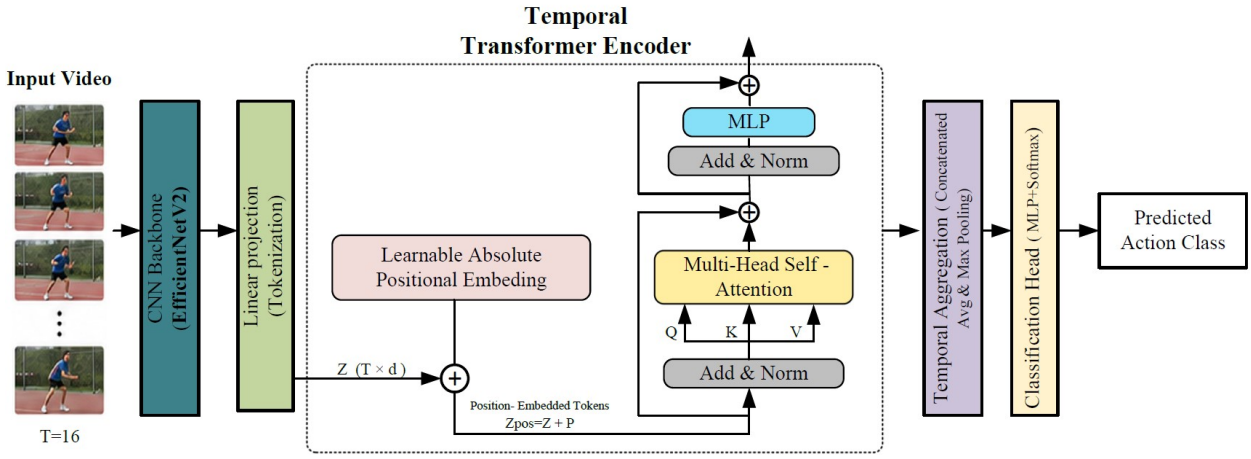


Figure 1. The framework of our baseline architecture with absolute positional embedding for HAR.

In a standard Vision Transformer, self-attention lacks an inherent notion of token order [13]. Therefore, our baseline model injects temporal order by adding a learnable absolute positional embedding $P \in \mathbb{R}^{T \times d}$ directly to the projected sequence Z:

$$Z_{pos} = Z + P. \tag{4}$$

The resulting position-embedded sequence Z_{pos} is then processed by multi-head self-attention (MHSA). In the first Transformer layer, Z_{pos} is linearly projected into queries (Q), keys (K), and values (V):

$$Q = Z_{pos}W_Q, \quad K = Z_{pos}W_K, \quad V = Z_{pos}W_V \tag{5}$$

The standard scaled dot-product attention is computed as [13]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_h}}\right)V \tag{6}$$

where d_h is the dimension of each attention head. Following the Transformer blocks, the frame-level representations are aggregated via a Temporal Aggregation module that concatenates average and max pooling results. The resulting video-level representation is then fed into an MLP classification head with Softmax activation to predict the action class. While this additive approach informs the model of absolute frame indices, it does not encode the relative temporal distance between two frames, which is relevant to action recognition.

3.4. RoPE-Enhanced Temporal Self-Attention

We replace our additive baseline embeddings with Rotary Position Embedding (RoPE) [33] to encode the temporal distance between frames. Figure 2 illustrates the overall architecture of the proposed model. Specifically, part (a) shows the macro-level pipeline, where sampled frames are sequentially processed to generate video-level predictions. To clarify the core of our temporal modeling, parts (c) and (d) detail the modified Multi-Head Self-Attention module and the scaled dot-product attention, respectively. Furthermore, the geometric interpretation of RoPE is presented in Part (b), demonstrating how the query (Q) and key (K) representations are rotated to encode relative temporal distances. The theoretical foundations and detailed mathematical formulations are elaborated below.

Unlike absolute encodings, RoPE injects positional information by rotating the query and key vectors in the multi-head attention module. For a frame token at temporal position t, RoPE applies a rotation matrix $R_{\theta,t}$ to the paired feature dimensions of the query and key

vectors. For a two-dimensional subspace of the feature space, the rotation transformation is defined as [33]:

$$R(t,\omega) = \begin{pmatrix} \cos(t\omega) & -\sin(t\omega) \\ \sin(t\omega) & \cos(t\omega) \end{pmatrix} \tag{7}$$

where ω is the frequency. In the proposed model, the rotated queries and keys are formulated as:

$$q'_t = R_{\theta,t} q_t, \quad k'_t = R_{\theta,t} k_t \tag{8}$$

The advantage of RoPE is that the dot-product similarity between a rotated query at temporal frame m and a rotated key at temporal frame n becomes strictly dependent on their relative temporal displacement $(n-m)$ [33]:

$$(q'_m)^T k'_n = q_m^T R_{\theta,n-m} k_n \tag{9}$$

This property makes attention scores sensitive to the relative temporal displacement between frames. The rotation frequencies assigned to the two-dimensional feature subspaces follow a geometric progression controlled by the rotary base parameter θ:

$$\omega_i = \theta^{-2i/d_h}, \qquad i = 0, 1, \ldots, \frac{d_h}{2} - 1 \tag{10}$$

where d is the Transformer embedding dimension, H is the number of attention heads, and $d_h = d/H$ denotes the feature dimension of each attention head. The head dimension d_h is required to be even. The index i identifies the adjacent feature pair $(2i, 2i+1)$. At temporal position t, each feature pair is rotated by the angle $t\omega_i$. In our implementation, RoPE is applied to all dimensions of the query and key vectors in each attention head, while the value vectors remain unchanged.

While the rotary base is commonly set to $\theta = 10,000$ in language models [33], its optimal value may depend on the temporal characteristics of the input sequence. Recent video-specific RoPE studies, such as VideoRoPE [35], have also examined frequency allocation in video representations, though typically in the context of large-scale multimodal models. In contrast, our controlled setup isolates the effect of frequency tuning in a lightweight RGB-only framework. According to Equation (10), a smaller value of θ produces faster positional rotations for feature pairs with $i > 0$, while a larger value produces slower positional variations. Therefore, we evaluate several values of θ in our ablation study to examine its effect on action recognition performance.

3.5. Classification Head and Training Objective

Following the L-layer RoPE-enhanced Transformer encoder, the output is a contextualized sequence of frame representations:

$$H = [h_1, h_2, \ldots, h_T] \in \mathbb{R}^{T \times d} \tag{11}$$

We apply both global average pooling and global max pooling across the temporal dimension. The average-pooled representation is computed as:

$$h_{avg} = \frac{1}{T} \sum_{t=1}^{T} h_t \in \mathbb{R}^d \tag{12}$$

The max-pooled representation is computed by taking the element-wise maximum across all frame representations:

$$h_{max} = \max_{1 \le t \le T} h_t \in \mathbb{R}^d \tag{13}$$

The two pooled representations are then concatenated to obtain the final video-level representation:

$$h_{video} = \text{Concat}(h_{avg}, h_{max}) \in \mathbb{R}^{2d} \tag{14}$$

where Concat$(\cdot)$ denotes vector concatenation. The resulting video-level representation is passed to a fully connected classification head to produce the final action probability distribution:

$$\hat{y} = \text{softmax}(W_c h_{video} + b_c) \tag{15}$$

All trainable components downstream of the frozen CNN feature extractor are optimized jointly using the standard categorical cross-entropy loss:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{16}$$

where C is the total number of action classes, y_c is the ground-truth label indicator, and $\hat{y}_c$ is the predicted probability for class c.

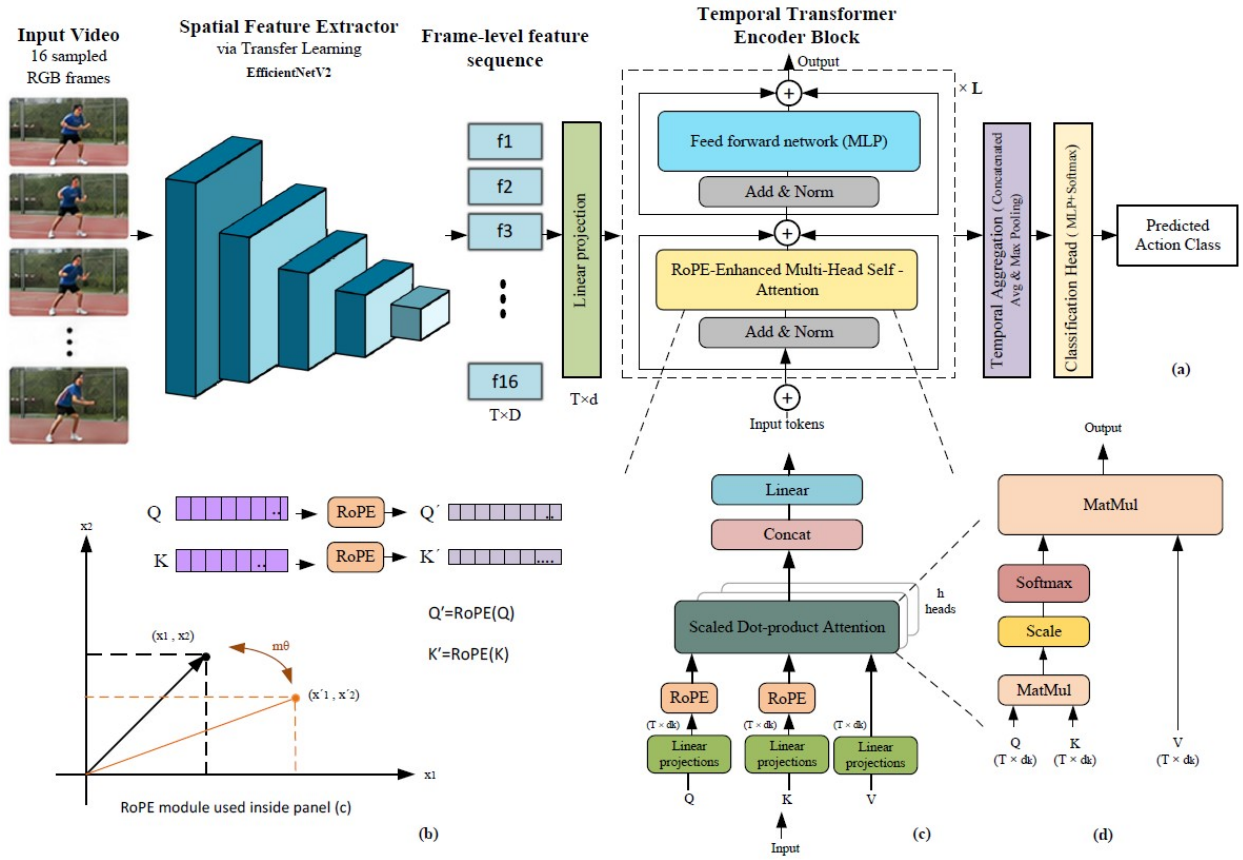


Figure 2. Proposed RoPE-enhanced CNN–Transformer architecture: (**a**) overall pipeline with a CNN spatial feature extractor, temporal Transformer encoder, temporal aggregation, and classification head; (**b**) geometric interpretation of RoPE applied to Q and K; (**c**) multi-Head Self-Attention module highlighting RoPE integration; and (**d**) scaled dot-product attention.

4. Experimental Setup

4.1. Datasets

We conduct experiments on two widely adopted human action recognition benchmarks, UCF101 and HMDB51, to evaluate the proposed RoPE-enhanced temporal Transformer.

- UCF101 [57]: This dataset consists of 13,320 video clips categorized into 101 action classes collected from YouTube. It presents significant challenges due to variations in camera motion, viewpoint, background clutter, and illumination.
- HMDB51 [58]: Comprising 6766 video clips across 51 action classes, HMDB51 is sourced from movies and web videos. It is generally considered more challenging

than UCF101 due to complex intra-class variations and a smaller number of training samples per category.

For both datasets, we follow the standard evaluation protocol using the three official train/test splits and report the average classification accuracy. Each UCF101 split contains approximately 9500 training and 3700 test videos. For validation, 15% of the official training set in each split was selected using stratified random sampling (Split 1: 8106 training, 1431 validation, and 3783 test; Split 2: 8148 training, 1,438 validation, and 3734 test; Split 3: 8180 training, 1444 validation, and 3696 test). Each HMDB51 split contains 3570 training and 1530 test videos before validation partitioning (3,034 training, 536 validation, and 1530 test videos per split). The official test sets were kept separate from training and validation.

4.2. Video Preprocessing and Data Augmentation

Effective preprocessing and augmentation are critical for training hybrid CNN-Transformer models, especially on relatively small datasets like HMDB51, to prevent overfitting and improve generalization. Given an input video, we uniformly sample $T = 16$ frames spanning the entire duration of the clip, a standard choice that balances temporal coverage with computational efficiency. For videos containing fewer than 16 frames, repeated frame indices are used to obtain the required sequence length. Each sampled frame is resized so that its shorter side is 160 pixels while preserving the aspect ratio. RGB Pixel values are supplied in $[0, 255]$, with normalization handled internally by EfficientNetV2B0.

For training videos, we apply a standard spatial data augmentation pipeline to the sampled frames. This includes random square cropping, with the crop area sampled uniformly between 80% and 100% of a 160×160 reference square, followed by resizing to 160×160 pixels. We also apply horizontal flipping with a probability of 0.5 and contrast jittering with a factor sampled uniformly from $[0.8, 1.2]$. The same augmentation parameters are shared across all frames of each video to preserve temporal coherence. Augmentation is applied only during training. During inference, we apply a center crop of 160×160 pixels to each resized frame without random augmentation.

4.3. Implementation Details

The proposed framework is implemented using a modular CNN-Transformer architecture. For spatial feature extraction, we utilize an EfficientNetV2B0 backbone pretrained on ImageNet. The classification head is removed, and the network weights are frozen during training to serve as an off-the-shelf feature extractor. The CNN extracts a feature vector of dimension $D = 1280$ for each sampled frame, which is subsequently projected to the Transformer's embedding dimension of $d = 512$ using a linear dense layer.

The Transformer uses a feed-forward dimension of 1024 with GELU activation and follows a Pre-LayerNorm structure. Dropout is applied within the attention and feed-forward sublayers and in the classification head. Training is performed for a maximum of 160 epochs.

The temporal sequence modeling is handled by a Transformer encoder composed of $L = 3$ layers, where each layer uses multi-head self-attention with $H = 8$ heads. To mitigate overfitting, a dropout rate of 0.2 is applied. In the proposed model, temporal order is explicitly encoded by applying Rotary Position Embedding (RoPE) to the query and key projections within the self-attention module.

During training, the network is optimized using the AdamW optimizer with categorical cross-entropy loss. A cosine decay schedule and early stopping based on the validation loss are used to support stable convergence and reduce overfitting. All models were implemented in Python 3.13.8 using the TensorFlow 2.20.0 deep learning framework and its integrated Keras API. The experiments were conducted using a machine equipped with an

NVIDIA GeForce RTX 2080 GPU with 8 GB of memory. A summary of the experimental hyperparameters and model configurations is provided in Table 1.

Table 1. Summary of the experimental hyperparameters and model configurations.

Hyperparameter/Setting	Value
Spatial Feature Extractor	EfficientNetV2B0
Input Frame Resolution	160×160
Number of Sampled Frames (T)	16
Transformer Hidden Dimension (d)	512
Number of Attention Heads (H)	8
Optimizer	AdamW
Base Learning Rate	1×10^{-4}
Learning Rate Schedule	Cosine decay
Weight Decay	1×10^{-4}
Batch Size	32

4.4. Evaluation Metric

The evaluation metric used in all experiments is Top-1 classification accuracy, defined as the percentage of correctly classified test videos:

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \times 100 \tag{17}$$

where N_{correct} represents the number of videos accurately assigned to their ground-truth action class, and N_{total} is the total number of evaluated videos.

4.5. Comparative Analysis: Baseline vs. Proposed RoPE Architecture

In this section, we examine whether modeling the relative positions among frames improves action recognition compared with learnable absolute positional embeddings. We first compare our baseline model with the proposed RoPE-enhanced model using the same architecture and training settings. Our evaluation focuses specifically on these two positional encoding approaches rather than a broader comparison of different positional encoding methods.

Both models use the same pretrained EfficientNetV2B0 as spatial feature extractor and process the same uniformly sampled sequence of $T = 16$ frames. They also use the same temporal Transformer configuration ($L = 3$, $H = 8$, and dropout $= 0.2$) and categorical cross-entropy loss. Our baseline model injects temporal order by adding learnable absolute positional embeddings to the frame tokens before self-attention. In contrast, the RoPE-based model omits these additive embeddings and instead integrates RoPE with a base parameter of $\theta = 10,000$ directly to the query and key projections within the multi-head self-attention module.

The results of this comparison are reported in Table 2. Under identical architectural and training settings, the RoPE-based model achieves modest improvements on both datasets. On UCF101, the accuracy increases from 91.51% to 91.59%, corresponding to an improvement of 0.08 percentage points. On HMDB51, the accuracy increases from 74.51% to 75.10%, an improvement of 0.59 percentage points. Because the architecture and training settings remain unchanged, the observed differences can be associated with the change in positional encoding under this configuration. Applying RoPE to the query and key projections allows the attention scores to reflect the relative positions among sampled frames. This is relevant to action recognition because many actions depend on the order and progression of motion across frames rather than the absolute position of an individual frame.

We also measured the inference time of the temporal module for both models using the same hardware setup and a batch size of 32. The reported values represent amortized processing time per video, excluding CNN feature extraction and video preprocessing. On UCF101, the baseline and RoPE-enhanced models record inference times of 0.50 ms/video and 0.36 ms/video, respectively. On HMDB51, the corresponding times are 0.36 ms/video and 0.44 ms/video. Overall, both models show comparable inference times, with no consistent speed advantage for either method. These measurements provide additional information about the computational performance of the temporal module under the evaluated experimental settings.

Table 2. Accuracy and temporal time of the baseline and RoPE-enhanced models under the same settings.

Model	UCF101		HMDB51	
	Acc (%)	Temporal Time (ms/Video)	Acc (%)	Temporal Time (ms/Video)
Baseline (ours)	91.51	0.50	74.51	0.36
RoPE (ours)	91.59	0.36	75.10	0.44

4.6. Ablation Study

The controlled comparison in the previous section evaluates the effect of replacing learnable absolute positional embeddings with RoPE under identical architectural and training settings. The following ablation studies focus on the proposed RoPE-enhanced model and investigate how Transformer depth, dropout, and the rotary base parameter affect its classification performance. The objective is to examine the sensitivity of the proposed model to these parameters and identify better-performing configurations within the evaluated settings. A sequential analysis was adopted to better understand how changes in these parameters affect model performance without searching for the optimal combination across all possible settings.

4.6.1. Model Capacity and Regularization

Our baseline and RoPE models compared earlier use the same configuration ($H = 8$, $L = 3$, and dropout $= 0.2$). This configuration is used as the reference for the following ablation experiments. This subsection examines the effects of Transformer depth and dropout on the RoPE-enhanced model for UCF101 and HMDB51. In all experiments, the number of attention heads and the rotary base parameter are fixed at $H = 8$ and $\theta = 10{,}000$, respectively. The hyperparameter analysis was conducted sequentially. Starting from the reference configuration, individual parameters were adjusted based on the validation performance observed in the preceding experiments. The official test sets were used only for final evaluation and were not involved in hyperparameter selection. The accuracies reported in Tables 3–6 represent the average test accuracy across the three official splits.

For UCF101, we first increase the Transformer depth from $L = 3$ to $L = 4$ while keeping dropout $= 0.2$. We then increase the dropout rate to 0.3 while keeping the four-layer architecture fixed. The results are presented in Table 3.

Table 3. Effect of Transformer depth and dropout on UCF101 ($\theta = 10,000$).

Model Variant	Heads	Layers (L)	Dropout	Acc (%)
RoPE	8	4	0.2	90.02
RoPE	8	4	0.3	91.41

As shown in Table 3, increasing the Transformer depth to four layers while keeping dropout = 0.2 results in an accuracy of 90.02%. This result indicates that increasing model capacity alone does not improve performance on UCF101. When the dropout rate is increased from 0.2 to 0.3, the accuracy improves by 1.39 percentage points, reaching 91.41%. This improvement suggests that stronger regularization helps the deeper model generalize more effectively. However, this result remains slightly below the 91.59% accuracy obtained using the reference configuration ($L = 3$, dropout = 0.2). Therefore, at $\theta = 10{,}000$, increasing the Transformer depth does not provide an additional improvement on UCF101.

For HMDB51, we first compare models with $L = 4$ and $L = 2$ at a fixed dropout rate of 0.3. Based on the improved validation performance of the lower-capacity configuration, we then keep the two-layer architecture fixed and progressively increase the dropout rate from 0.3 to 0.5. The results are reported in Table 4.

Table 4. Effect of Transformer depth and dropout on HMDB51 ($\theta = 10,000$).

Model Variant	Heads	Layers (L)	Dropout	Acc (%)
RoPE	8	4	0.3	72.68
RoPE	8	2	0.3	74.90
RoPE	8	2	0.4	76.47
RoPE	8	2	0.5	78.82

As shown in Table 4, reducing the Transformer depth from four to two layers at dropout = 0.3 improves the accuracy from 72.68% to 74.90%. This result suggests that the lower-capacity model is more suitable for HMDB51. With the number of layers fixed at $L = 2$, increasing the dropout rate to 0.4 and 0.5 further improves the accuracy to 76.47% and 78.82%, respectively. The best result is therefore obtained with $L = 2$ and dropout = 0.5, representing a 3.72-percentage-point improvement over the reference configuration. This suggests that for HMDB51, a shallower Transformer with stronger regularization is preferable. Overall, the results show that the most suitable balance between model capacity and regularization differs across the two datasets.

4.6.2. Impact of the Rotary Base Parameter (θ)

A key component of the RoPE mechanism is the base parameter, θ, which directly controls the rotation frequency within the feature space. A smaller base value induces a rapid phase change across the sequence, whereas a larger base results in a smoother, more gradual rotation. In natural language processing, setting $\theta = 10{,}000$ is the standard practice. However, video sequences exhibit different characteristics; human motion across sampled frames evolves continuously and gradually. We hypothesized that a larger base value might provide a more suitable inductive bias for video sequences, allowing temporal positions to be integrated more smoothly into the attention computation.

The range of θ values was selected separately for each dataset based on the results of the preceding experiments. Table 5 illustrates the effect of varying θ on the UCF101 dataset. Increasing the base from the standard 10,000 to 50,000 improves the accuracy from 91.41% to 91.91%. This suggests that a larger rotary base parameter can improve classification accuracy under the evaluated configuration. This behavior may be related to the effect of θ on the frequency of positional encoding. A smaller θ produces faster changes in the rotary representation, while a larger θ results in smoother positional variations across sampled frames.

Table 5. Comparison of different rotary base parameters (θ) and model configurations on the UCF101 dataset.

Rotary Base ($\boldsymbol{\theta}$)	Model Configuration	Acc (%)
1000	$L = 4$, Dropout = 0.3	90.30
5000	$L = 4$, Dropout = 0.3	91.48
10,000	$L = 4$, Dropout = 0.3	91.41
50,000	$L = 4$, Dropout = 0.3	91.91

As shown in Table 6, we evaluate the effect of the rotary base parameter on HMDB51 while keeping the model configuration fixed at $L = 2$ and dropout = 0.5. The highest accuracy of 78.82% is obtained with $\theta = 10,000$. Using either a smaller value ($\theta = 5000$) or larger values ($\theta = 20,000$ and $\theta = 50,000$) reduces the accuracy to 76.80%, 77.65%, and 75.03%, respectively. In particular, increasing the rotary base from 10,000 to 50,000 decreases the accuracy by 3.79 percentage points. The different trends on UCF101 and HMDB51 show that increasing θ does not always improve accuracy. As described in Equation (10), a larger θ leads to smaller positional phase changes across sampled frames. HMDB51 contains fewer videos and greater intra-class variation than UCF101, which may partly explain the different results. Further feature-level analysis is needed to explain these differences more precisely.

Table 6. Comparison of rotary base parameters (θ) and model configurations on HMDB51.

Rotary Base ($\boldsymbol{\theta}$)	Model Configuration	Acc (%)
5000	$L = 2$, Dropout = 0.5	76.80
10000	$L = 2$, Dropout = 0.5	78.82
20,000	$L = 2$, Dropout = 0.5	77.65
50,000	$L = 2$, Dropout = 0.5	75.03

4.7. Comparison with State-of-the-Art Methods

We compare our RoPE-enhanced framework with several representative action recognition models on the UCF101 and HMDB51 datasets to evaluate the effectiveness of the proposed architecture. As summarized in Table 7, the comparison includes only methods that use 2D spatial feature extractors pretrained on ImageNet. We exclude models based on 3D convolutions, along with those pretrained on large video datasets such as Kinetics-400, because their computational and training requirements differ.

Our results in Table 7 correspond to the best configurations identified for UCF101 and HMDB51 in the ablation studies presented in Tables 5 and 6, respectively.

The experimental results demonstrate that the proposed RoPE-enhanced framework achieves competitive performance on HMDB51, with an accuracy of 78.82%. It outperforms AST-VLAD, TSN + AS, STS, and TTSN, with gains ranging from 5.72 percentage points over AST-VLAD to 18.62 percentage points over TTSN. Among the evaluated methods, *k*-NN ViViT reports a higher accuracy of 82.50% using a ViT-Base backbone, while our framework uses EfficientNetV2 as its spatial feature extractor.

In addition, our proposed model achieves an accuracy of 91.91% on UCF101, surpassing the TTSN, STS, and STDAN models by 5.51, 1.81, and 0.91 percentage points, respectively. Although some methods based on heavier spatial backbones report slightly higher accuracies on UCF101, our model remains competitive while using a lightweight EfficientNetV2 backbone. Overall, these results suggest that incorporating relative positional information into temporal self-attention is a promising approach for modeling video dynamics within a lightweight CNN–Transformer framework.

Table 7. Comparison with representative state-of-the-art action recognition methods on UCF101 and HMDB51 datasets. Values in bold represent our proposed method's achievements.

Method	Year	Spatial Extractor/Pretraining	UCF101 (%)	HMDB51 (%)
STS [59]	2020	ResNet (ImageNet)	90.10	62.40
STDAN [60]	2020	VGGNet (ImageNet)	91.00	60.40
TSN + AS [61]	2022	BN-Inception (ImageNet)	94.50	71.60
TTSN [62]	2023	ResNet-50 (ImageNet)	86.40	60.20
AST-VLAD [32]	2024	BN-Inception (ImageNet)	96.00	73.10
k-NN ViViT [46]	2024	ViT-Base (ImageNet)	94.20	82.50
Baseline (Ours)	**-**	**EfficientNetV2B0 (ImageNet)**	91.51	74.51
Tuned RoPE-T (Ours)	**-**	**EfficientNetV2B0 (ImageNet)**	**91.91**	**78.82**

4.8. Qualitative Analysis and Visualization

In this section, we visualize temporal attention scores alongside spatial class activation maps (Grad-CAM) to better understand how the proposed architecture processes spatial and temporal information. For this qualitative analysis, we use our best-performing RoPE-enhanced models for each dataset, which achieved accuracies of 91.91% on UCF101 and 78.82% on HMDB51. Each visualization presents the temporal attention distribution across the full 16-frame sequence in part (a). The four frames with the highest temporal attention scores are shown as original RGB frames in part (b), together with their corresponding Grad-CAM spatial overlays in part (c).

Figure 3 visualizes the network's behavior on the "TrampolineJumping" action sequence from the UCF101 dataset. This action was selected due to its inherently high spatial displacement. The temporal attention weights demonstrate that the RoPE mechanism identifies the most discriminative temporal moments of the action. Spatially, the Grad-CAM overlays reveal that the model accurately tracks the human subject in mid-air. The high-activation regions, shown in red, are concentrated around the person, while the static background receives little attention.

Similarly, Figure 4 presents the visualization for the "kick_ball" action sequence from the more challenging HMDB51 dataset. This example illustrates the model's capacity to comprehend complex human–object interactions in a dynamic environment. The temporal attention scores accurately highlight the critical frames during the ball's impact. More importantly, the Grad-CAM spatial overlays demonstrate exceptional fine-grained localization; the highest-activation regions (red) are concentrated on the player's moving leg and the ball itself.

These visualizations provide a qualitative view of how the RoPE-enhanced model distributes attention across sampled frames and spatial regions. Together with the classification results, they illustrate the behavior of the proposed temporal encoding mechanism. However, attention visualizations and classification accuracy alone do not directly establish improved temporal reasoning, which would require additional evaluations specifically designed to measure sensitivity to temporal order.

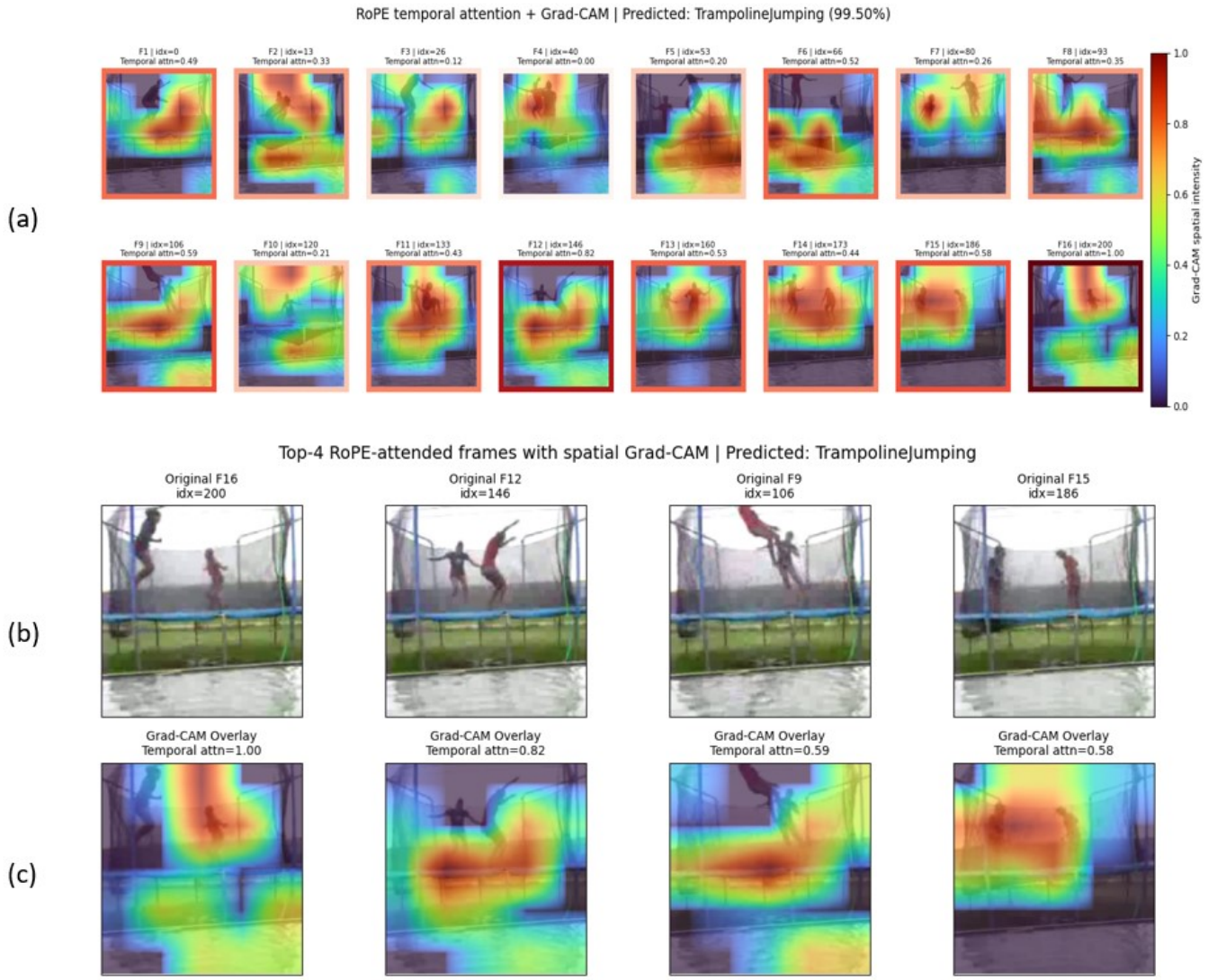


Figure 3. Qualitative visualization of the proposed RoPE-enhanced model on the UCF101 dataset (action class: *TrampolineJumping*). (**a**) Temporal attention and Grad-CAM across the 16-frame sequence. (**b**) Original RGB frames of the top-4 attended temporal moments. (**c**) Grad-CAM overlays demonstrating precise spatial tracking of the bounding subject.

We also examined two misclassified examples from the predictions of our RoPE-enhanced model on UCF101 to better understand its limitations. In the first example, a video from the Nunchucks class was incorrectly classified as PlayingDaf, while in the second, a Fencing video was classified as LongJump. These misclassifications may be related to similarities in body movements, variations in viewpoint, or the limited temporal information captured by the 16 sampled frames.

These visualizations suggest that the model focuses on informative temporal transitions and relevant spatial regions while giving less attention to the grassy background and less relevant frames. Together, they complement the quantitative results by showing the temporal and spatial regions emphasized by the RoPE-enhanced model.

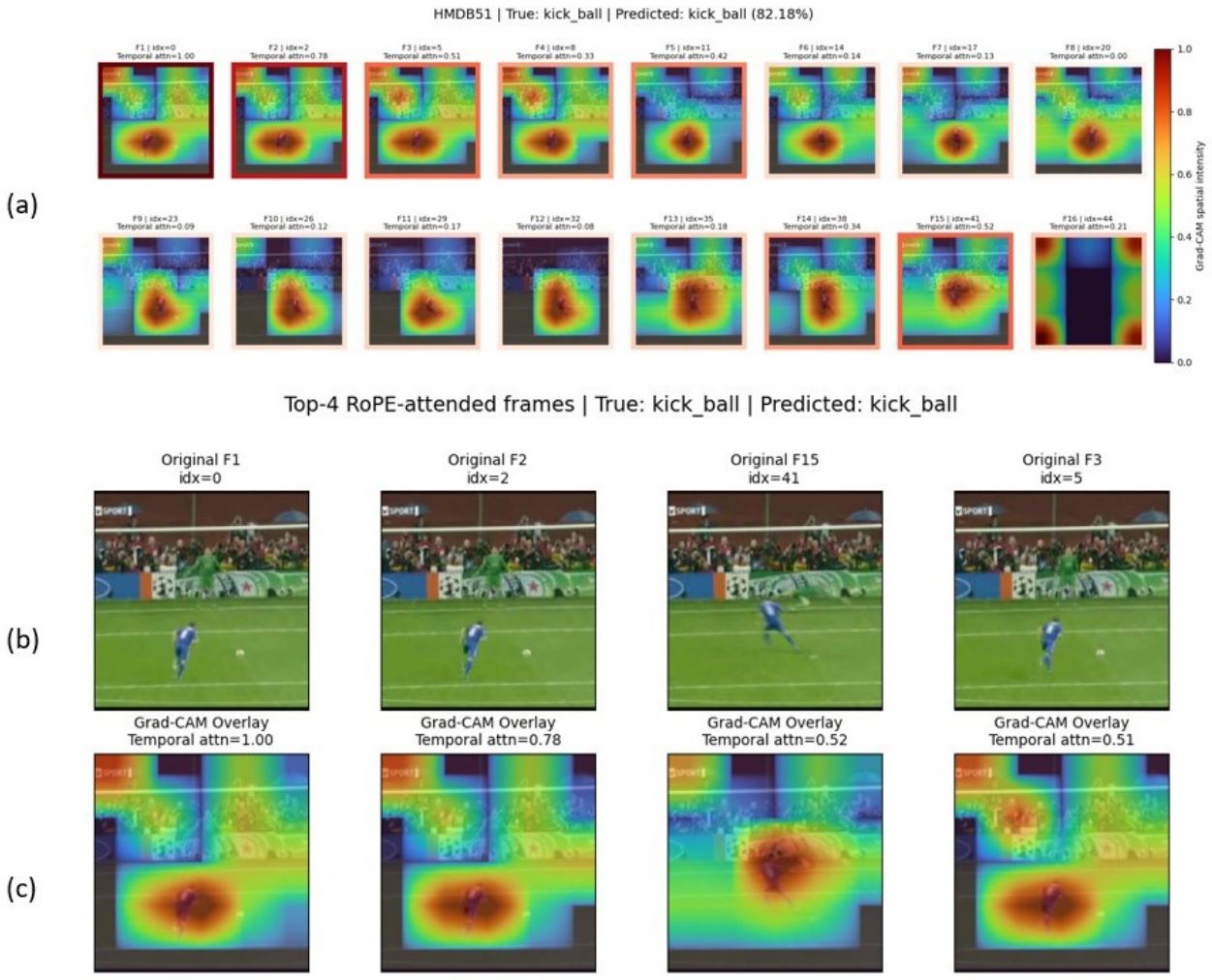


Figure 4. Qualitative visualization of the proposed RoPE-enhanced model on the HMDB51 dataset (action class: *kick_ball*). (**a**) Temporal attention and Grad-CAM across the 16-frame sequence. (**b**) Original RGB frames of the top-4 attended temporal moments. (**c**) Grad-CAM overlays highlighting accurate spatial localization of the human–object interaction.

5. Conclusions and Future Work

In this paper, we proposed RoPE-T, a RoPE-enhanced temporal Transformer model for video-based action recognition, which incorporates relative positional information among sampled frames and improves recognition accuracy compared with our baseline model. While the controlled comparison shows modest improvements from RoPE, the final performance also reflects the effects of tuning Transformer depth, dropout, and the rotary base parameter. Our experiments show that a larger rotary base parameter (θ) achieves the highest accuracy on UCF101, whereas the default value performs best on HMDB51. This highlights the dataset-dependent effect of the rotary base parameter on action recognition performance. Experimental results demonstrate that the proposed framework achieves higher accuracy than several representative action recognition methods on both benchmarks. However, our positional encoding comparison is limited to learnable absolute embeddings and does not include other positional encoding methods. Using a frozen spatial backbone and 16 sampled frames may limit the recognition of fine-grained and long-duration actions. Our future work will focus on extending the RoPE-enhanced temporal encoding to multimodal settings and comparing RoPE-T with other positional encoding strategies. We will

also explore end-to-end spatial fine-tuning and denser temporal sampling to better capture fine-grained motion and long-term temporal dependencies.

Author Contributions: Conceptualization, F.S. and M.T.S.; formal analysis, F.S.; investigation, F.S.; supervision, M.T.S.; writing—original draft preparation, F.S.; writing—review and editing, F.S. and M.T.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: The UCF101 and HMDB51 datasets can be found at https://www.crcv.ucf.edu/data/UCF101.php (accessed on 15 October 2026) and https://serre.lab.brown.edu/hmdb51.html (accessed on 15 October 2026), respectively.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sun, Z.; Ke, Q.; Rahmani, H.; Bennamoun, M.; Wang, G.; Liu, J. Human action recognition from various data modalities: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 3200–3225.
2. Sultani, W.; Chen, C.; Shah, M. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6479–6488.
3. Gao, Y.; Xiang, X.; Xiong, N.; Huang, B.; Lee, H.J.; Alrifai, R.; Jiang, X.; Fang, Z. Human action monitoring for healthcare based on deep learning. *IEEE Access* **2018**, *6*, 52277–52285. [CrossRef]
4. Thomas, G.; Gade, R.; Moeslund, T.B.; Carr, P.; Hilton, A. Computer vision for sports: Current applications and research topics. *Comput. Vis. Image Underst.* **2017**, *159*, 3–18. [CrossRef]
5. Shotton, J.; Sharp, T.; Kipman, A.; Fitzgibbon, A.; Finocchio, M.; Blake, A.; Cook, M.; Moore, R. Real-time human pose recognition in parts from single depth images. *Commun. ACM* **2013**, *56*, 116–124. [CrossRef]
6. Pan, Y.; Mei, T.; Yao, T.; Li, H.; Rui, Y. Jointly modeling embedding and translation to bridge video and language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4594–4602.
7. Beddiar, D.R.; Nini, B.; Sabokrou, M.; Hadid, A. Vision-based human activity recognition: A survey. *Multimed. Tools Appl.* **2020**, *79*, 30509–30555. [CrossRef]
8. Shafizadegan, F.; Naghsh-Nilchi, A.R.; Shabaninia, E. Multimodal vision-based human action recognition using deep learning: A review. *Artif. Intell. Rev.* **2024**, *57*, 178. [CrossRef]
9. Surek, G.A.S.; Seman, L.O.; Stefenon, S.F.; Mariani, V.C.; Coelho, L.D.S. Video-based human activity recognition using deep learning approaches. *Sensors* **2023**, *23*, 6384. [CrossRef] [PubMed]
10. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 8–13 December 2014; pp. 568–576.
11. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
12. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the Kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
14. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. ViViT: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 6836–6846.
15. Alomar, K.; Aysel, H.I.; Cai, X. CNNs, RNNs and transformers in human action recognition: A survey and a hybrid model. *Artif. Intell. Rev.* **2025**, *58*, 387. [CrossRef]
16. Stergiou, A.; Poppe, R. About time: Advances, challenges, and outlooks of action understanding. *Int. J. Comput. Vis.* **2025**, *133*, 6251–6315. [CrossRef]
17. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2740–2755. [CrossRef] [PubMed]
18. Li, Q.; Yang, W.; Chen, X.; Yuan, T.; Wang, Y. Temporal segment connection network for action recognition. *IEEE Access* **2020**, *8*, 179118–179127. [CrossRef]

19. Donahue, J.; Hendricks, L.A.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
20. Varol, G.; Laptev, I.; Schmid, C. Long-term temporal convolutions for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1510–1517. [CrossRef] [PubMed]
21. Li, D.; Yao, T.; Duan, L.Y.; Mei, T.; Rui, Y. Unified spatio-temporal attention networks for action recognition in videos. *IEEE Trans. Multimed.* **2018**, *21*, 416–428. [CrossRef]
22. Du, W.; Wang, Y.; Qiao, Y. Recurrent spatial-temporal attention network for action recognition in videos. *IEEE Trans. Image Process.* **2017**, *27*, 1347–1360. [CrossRef] [PubMed]
23. Peng, Y.; Zhao, Y.; Zhang, J. Two-stream collaborative learning with spatial-temporal attention for video classification. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 773–786. [CrossRef]
24. Yang, G.; Yang, Y.; Lu, Z.; Yang, J.; Liu, D.; Zhou, C.; Fan, Z. STA-TSN: Spatial-temporal attention temporal segment network for action recognition in video. *PLoS ONE* **2022**, *17*, e0265115. [CrossRef] [PubMed]
25. Ding, H.; Guo, C.; Sun, J.; Jiang, X.; Shi, H.; Li, J. Motion-driven adaptive frame selection strategy for video action recognition. *EURASIP J. Image Video Process.* **2025**, *2025*, 12. [CrossRef]
26. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021.
27. Shi, C.; Liu, S. Human action recognition with transformer based on convolutional features. *Intell. Decis. Technol.* **2024**, *18*, 881–896. [CrossRef]
28. Hussain, A.; Khan, S.U.; Khan, N.; Bhatt, M.W.; Farouk, A.; Bhola, J.; Baik, S.W. A hybrid transformer framework for efficient activity recognition using consumer electronics. *IEEE Trans. Consum. Electron.* **2024**, *70*, 6800–6807. [CrossRef]
29. Zhou, A.; Ma, Y.; Ji, W.; Zong, M.; Yang, P.; Wu, M.; Liu, M. Multi-head attention-based two-stream EfficientNet for action recognition. *Multimed. Syst.* **2023**, *29*, 487–498. [CrossRef]
30. Chen, T.; Mo, L. Swin-Fusion: Swin-Transformer with feature fusion for human action recognition. *Neural Process. Lett.* **2023**, *55*, 11109–11130. [CrossRef]
31. Li, X.; Yang, W.; Wang, K.; Wang, T.; Zhang, C. MANet: Motion-aware network for video action recognition. *Complex Intell. Syst.* **2025**, *11*, 167. [CrossRef]
32. Weng, Z.; Li, X.; Xiong, S. Action recognition using attention-based spatio-temporal VLAD networks and adaptive video sequences optimization. *Sci. Rep.* **2024**, *14*, 26202. [CrossRef] [PubMed]
33. Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; Liu, Y. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing* **2024**, *568*, 127063. [CrossRef]
34. Neimark, D.; Bar, O.; Zohar, M.; Asselmann, D. Video transformer network. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, QC, Canada, 11–17 October 2021; pp. 3163–3172.
35. Wei, X.; Liu, X.; Zang, Y.; Dong, X.; Zhang, P.; Cao, Y.; Tong, J.; Duan, H.; Guo, Q.; Wang, J.; et al. VideoRoPE: What makes for good video rotary position embedding? In Proceedings of the 42nd International Conference on Machine Learning (ICML), Vancouver, BC, Canada, 13–19 July 2025; Volume 267, pp. 66118–66136.
36. Liu, Z.; Guo, L.; Tang, Y.; Yue, T.; Cai, J.; Ma, K.; Liu, Q.; Chen, X.; Liu, J. VRoPE: Rotary position embedding for video large language models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), Suzhou, China, 4–9 November 2025; pp. 14460–14472.
37. Ma, Y.; Wang, R.; Zong, M.; Ji, W.; Wang, Y.; Ye, B. Convolutional transformer network for fine-grained action recognition. *Neurocomputing* **2024**, *569*, 127027. [CrossRef]
38. 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] [PubMed]
39. Nguyen, H.P.; Ribeiro, B. Video action recognition collaborative learning with dynamics via PSO-ConvNet Transformer. *Sci. Rep.* **2023**, *13*, 14624. [CrossRef] [PubMed]
40. Zhang, H.; Hu, Z.; Yu, D.; Guan, L.; Liu, X.; Ma, C. Multipath attention and adaptive gating network for video action recognition. *Neural Process. Lett.* **2024**, *56*, 124. [CrossRef]
41. Wu, H.; Ma, X.; Li, X.; Song, R. GLFormer: A hierarchical cross-layer transformer decoding framework with global–local feature collaboration for video action recognition. *Complex Intell. Syst.* **2026**, *12*, 171. [CrossRef]
42. Bertasius, G.; Wang, H.; Torresani, L. Is space-time attention all you need for video understanding? In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual Event, 18–24 July 2021; Volume 139, pp. 813–824.
43. Selva, J.; Johansen, A.S.; Escalera, S.; Nasrollahi, K.; Moeslund, T.B.; Clapés, A. Video transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12922–12943. [CrossRef] [PubMed]

44. 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 (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 3202–3211.
45. 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 (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 4804–4814.
46. Sun, W.; Ma, Y.; Wang, R. k-NN attention-based video vision transformer for action recognition. *Neurocomputing* **2024**, *574*, 127256. [CrossRef]
47. Wang, J.; Peng, X.; Qiao, Y. Cascade multi-head attention networks for action recognition. *Comput. Vis. Image Underst.* **2020**, *192*, 102898. [CrossRef]
48. Wang, J.; Shao, Z.; Huang, X.; Lu, T.; Zhang, R.; Lv, X. Spatial–temporal pooling for action recognition in videos. *Neurocomputing* **2021**, *451*, 265–278. [CrossRef]
49. Li, J.; Wei, P.; Zheng, N. Nesting spatiotemporal attention networks for action recognition. *Neurocomputing* **2021**, *459*, 338–348. [CrossRef]
50. Nguyen, T.V.; Song, Z.; Yan, S. STAP: Spatial-temporal attention-aware pooling for action recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *25*, 77–86. [CrossRef]
51. Joudaki, M.; Imani, M.; Arabnia, H.R. TransMODAL: A dual-stream transformer with adaptive co-attention for efficient human action recognition. *Electronics* **2025**, *14*, 3326. [CrossRef]
52. Press, O.; Smith, N.A.; Lewis, M. Train short, test long: Attention with linear biases enables input length extrapolation. In Proceedings of the 10th International Conference on Learning Representations (ICLR), Virtual Event, 25–29 April 2022.
53. Ma, Y.; Wang, R. Relative-position embedding-based spatially and temporally decoupled transformer for action recognition. *Pattern Recognit.* **2024**, *145*, 109905. [CrossRef]
54. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-attention with relative position representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; Volume 2, pp. 464–468.
55. Heo, B.; Park, S.; Han, D.; Yun, S. Rotary position embedding for vision transformer. In Proceedings of the 18th European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 289–305.
56. Tan, M.; Le, Q.V. EfficientNetV2: Smaller Models and Faster Training. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual Event, 18–24 July 2021; Volume 139, pp. 10096–10106.
57. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human action classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
58. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.
59. Liu, Z.; Li, Z.; Wang, R.; Zong, M.; Ji, W. Spatiotemporal saliency-based multi-stream networks with attention-aware LSTM for action recognition. *Neural Comput. Appl.* **2020**, *32*, 14593–14602. [CrossRef]
60. Zhang, Z.; Lv, Z.; Gan, C.; Zhu, Q. Human action recognition using convolutional LSTM and fully connected LSTM with different attentions. *Neurocomputing* **2020**, *410*, 304–316. [CrossRef]
61. Dong, W.; Zhang, Z.; Song, C.; Tan, T. Identifying the key frames: An attention-aware sampling method for action recognition. *Pattern Recognit.* **2022**, *130*, 108797. [CrossRef]
62. Zhang, Y.; Li, J.; Jiang, N.; Wu, G.; Zhang, H.; Shi, Z.; Liu, Z.; Wu, Z.; Liu, X. Temporal transformer networks with self-supervision for action recognition. *IEEE Internet Things J.* **2023**, *10*, 12999–13011. [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.