Multi-Modal Video Action Recognition with Learnable Frame Pruning via Temporal Token Scoring

Takashi Higashi[1], Ryuto Ishibashi[2], Lin Meng[2]
{ri0117fe@ed.ritsumei.ac.jp[1],
ri0097fx@ed.ritsumei.ac.jp[2], menglin@fc.ritsumei.ac.jp[2]}
[1]Graduate School of Science and Engineering, Ritsumeikan University

[2]College of Science and Engineering, Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577, Japan

Abstract. This study proposes a Multi-Modal Action Density Scoring (MADS) framework that integrates frame pruning into a ViViT-based multi-modal video recognition model to improve computational efficiency while maintaining accuracy. MADS introduces two frame selection strategies: a Learnable Threshold and a Top-K Segment Method. The Learnable Threshold adaptively determines pruning levels via a learnable parameter guided by a statistical loss, whereas the Top-K Segment Method divides a video into temporal segments and selects the most informative frames based on normalized importance scores. Experiments on the NTU RGB+D dataset show that the Top-K Segment Method achieves up to 36% FLOPs reduction with only 0.5% accuracy drop, outperforming the Learnable Threshold. Qualitative analysis further confirms that Top-K Segment preserves temporally distributed and semantically rich frames, maintaining motion continuity and visual interpretability. These results highlight MADS as a flexible and efficient framework for real-time and resource-constrained video recognition.

Keywords: Video recognition, Action recognition, ViViT, Pruning

1 Introduction

In recent years, artificial intelligence (AI) has achieved remarkable progress, and correspondingly, video recognition tasks have rapidly advanced. AI-driven video recognition is expected to have broad applications across various domains such as healthcare, surveillance, and transportation, offering significant societal benefits. For instance, in healthcare and nursing, AI-based video recognition can assist in reducing the workload of professionals and provide real-time monitoring support. Furthermore, analyzing surveillance camera footage can help detect violence or illegal activities, thereby enhancing security. As such, AI-based video recognition is becoming increasingly important across multiple fields. However, deploying AI systems and services in real-world applications presents several challenges. Specifically, high recognition accuracy, computational efficiency,

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365275

model compactness, and real-time inference are required. To address these challenges, this study focuses on improving recognition accuracy through multi-modal learning and enhancing computational efficiency through frame pruning. Enhancing both recognition accuracy and computational efficiency is critical for enabling reliable and real-time video recognition systems applicable to a wide range of tasks.

While multi-modal analysis is effective for improving recognition accuracy, it can increase computational costs due to the larger amount of data. To overcome this issue, this study simultaneously performs frame pruning to remove redundant information, aiming to maintain high computational efficiency while achieving accurate recognition. Specifically, we employ a video vision transformer (ViViT) [1] model for multi-modal analysis and select important frames based on a scoring mechanism to reduce unnecessary frames. This strategy enables the elimination of redundant information in static frames or regions with minimal temporal changes, thereby reducing computational load and improving efficiency in video recognition.

The organization of this paper is as follows:

- Section 2 provides an overview of related research pertinent to this study.
- Sections 3 and 4 describe the proposed method, including frame reduction based on importance scoring, and the experimental setup.
- Section 5 presents the experimental results, evaluation, and discussion.
- Finally, Section 7 concludes this work and outlines future directions.

2 Related Work

2.1 Video Vision Transformer (ViViT)

In this study, we employ the Video Vision Transformer (ViViT) [1] as the backbone for video recognition. ViViT extends the Vision Transformer (ViT) [2] to handle video data by partitioning each frame into non-overlapping patches and embedding them as high-dimensional tokens, capturing both spatial and temporal information.

ViViT processes the token sequence through a hierarchical attention mechanism. Spatial attention is first applied within each frame to extract local spatial features. Subsequently, temporal attention integrates information across frames, modeling long-range dependencies throughout the video. This two-stage attention design allows ViViT to efficiently aggregate spatial and temporal features.

A notable characteristic of ViViT is the presence of many redundant tokens, particularly in regions with little motion or static backgrounds. Leveraging this property, frame pruning can be applied to remove less informative frames prior to attention computation, resulting in substantial reduction of temporal computation. Specifically, if the original video has T frames and only $T_r < T$ frames are selected, the FLOPs for temporal attention can be roughly reduced as:

$$\mathrm{FLOPs}_{\mathrm{temporal}} \approx \frac{T_r^2}{T^2} \cdot \mathrm{FLOPs}_{\mathrm{original}}. \tag{1}$$

Spatial attention is computed independently for each frame, so the reduction in FLOPs is linear:

$$\mathrm{FLOPs}_{\mathrm{spatial}} \approx \frac{T_r}{T} \cdot \mathrm{FLOPs}_{\mathrm{original}}. \tag{2}$$

In this way, ViViT, due to its characteristic redundancy in tokens, allows frame pruning to significantly improve temporal computation efficiency. This provides substantial advantages for real-time inference and resource-constrained environments. Moreover, ViViT's hierarchical attention mechanism and high-dimensional token embeddings enable it to preserve essential spatial and temporal information while reducing computational cost, minimizing the impact on recognition accuracy.

2.2 Pruning and Frame Pruning

Pruning is a technique to remove unnecessary components of a neural network, reducing computational cost and memory usage while maintaining performance [3]. Typically, parameters such as weights or channels are removed based on their contribution or structural importance.

In video recognition, temporal redundancy is common, and frame pruning has been proposed to remove less informative frames [4]. Frame pruning determines which frames to remove based on redundancy or relevance of the information contained. Representative conventional approaches include:

- Inter-frame similarity-based methods: For consecutive frames F_t, F_{t+1} with feature vectors x_t, x_{t+1}, frame F_{t+1} is removed if the similarity $\mathrm{sim}(x_t, x_{t+1})$ exceeds a threshold ε [5]:

$$F_{t+1} \text{ is pruned if } \mathrm{sim}(x_t, x_{t+1}) \geq \varepsilon. \tag{3}$$

- Segmentation-based methods: Frames are removed if the ratio of object pixels $|R_{\mathrm{object}}|$ to total pixels $|F_t|$ is below a threshold δ, where R_{object} is the segmented region of humans or objects [6]:

$$F_t \text{ is pruned if } \frac{|R_{\mathrm{object}}|}{|F_t|} < \delta. \tag{4}$$

ViViT, with its high-dimensional token representations and hierarchical attention mechanism, is particularly suited for frame pruning. Even after removing redundant frames, it can preserve essential spatial and temporal information, making it effective for efficient video recognition.

2.3 Multi-Modal Analysis

Multi-modal analysis integrates different data types, such as images, text, and audio, to extract information unavailable from a single modality [7].

Let $f_m \in \mathbb{R}^{d_m}$ be the feature vector from modality m. Typical fusion strategies include:

$$f_{\mathrm{fusion}} = \left[f_1, f_2, \ldots, f_M\right] \quad \text{(concatenation)}, \tag{5}$$

$$f_{\text{fusion}} = \sum_{m=1}^{M} w_m f_m, \quad \sum_{m=1}^{M} w_m = 1 \quad \text{(weighted sum).} \tag{6}$$

For each modality with T frames, N_m tokens/patches, and embedding dimension d_m, the computation is roughly:

$$\text{FLOPs}_m \sim T \cdot N_m \cdot d_m^2. \tag{7}$$

Fusing multiple modalities adds cross-modal attention or fusion cost C_{fusion}:

$$\text{FLOPs}_{\text{multi}} \sim \sum_{m=1}^{M} (T \cdot N_m \cdot d_m^2) + C_{\text{fusion}}. \tag{8}$$

ViViT splits frames into patches and applies spatial and temporal attention, requiring substantial computation even for a single modality:

$$\text{FLOPs}_{\text{ViViT}} \sim T \cdot N \cdot d^2 + T^2 \cdot d^2. \tag{9}$$

With multiple modalities, the computation grows roughly as T^2 and M^2, making real-time or resource-constrained deployment challenging.

3 Overview of the Proposal

In this work, we propose a Multi-modal Action Density Scoring (MADS) framework, which incorporates a frame importance selection mechanism into the ViViT model enabling multimodal analysis as introduced in our previous study [8]. MADS allows for efficient frame pruning in video-based action recognition.

3.1 Multi-modal Analysis and Attention Control

After extracting spatial features from RGB, Depth, and Skeleton modalities via ViViT (Fig. 1), we integrate the modalities using an Attention Mask (Fig. 2). Specifically, we allow interaction only between CLS tokens across modalities to prevent excessive interference while efficiently combining important information:

$$X_{\text{att}} = \text{AttentionMask}(CLS_{\text{RGB}}, CLS_{\text{Depth}}, CLS_{\text{Skeleton}}) \tag{10}$$

This approach preserves modality-specific characteristics while providing a rich multimodal representation.

3.2 Multi-Modal Action Density Scoring Module

For each modality, we compute a frame importance score from the CLS tokens:

$$\begin{aligned} s_f^m &= \text{MLP}_m(\text{CLS}_f^m), \\ &\quad m \in \{\text{RGB, Depth, Skeleton}\}, \quad f \in [1, F] \end{aligned} \tag{11}$$

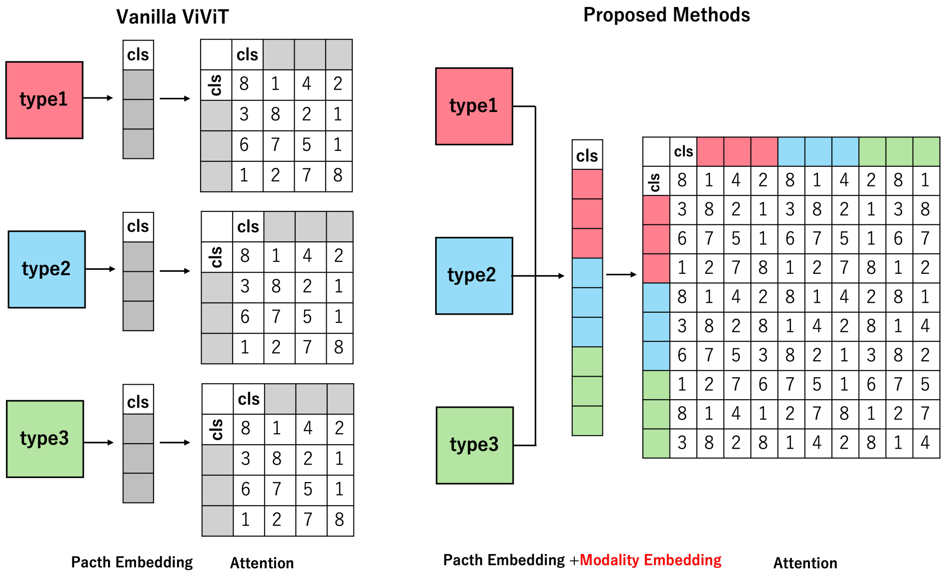

Fig. 1. Overview Diagram of Multi-modal Analysis [8].

The final frame score S_f is defined as the sum of the modality-specific scores:

$$S_f = \sum_m s_f^m \tag{12}$$

We propose two strategies for frame selection.

3.2.1 Learnable Threshold

We introduce a learnable threshold for frame pruning. For each frame f and modality m, the importance score s_f^m is computed, and pruning is controlled based on the distribution of these scores.

First, the mean μ and standard deviation σ of the scores across frames are calculated as:

$$\mu = \frac{1}{F} \sum_{f=1}^{F} s_f^m, \quad \sigma = \sqrt{\frac{1}{F} \sum_{f=1}^{F} (s_f^m - \mu)^2}. \tag{13}$$

Next, the target z value corresponding to the desired pruning rate r is computed using the inverse cumulative distribution function (CDF) of the standard normal distribution Φ^{-1}:

$$z_r = \Phi^{-1}(r), \tag{14}$$

where z_r represents the threshold that removes the top r proportion of frames after standardization (Fig. 3).

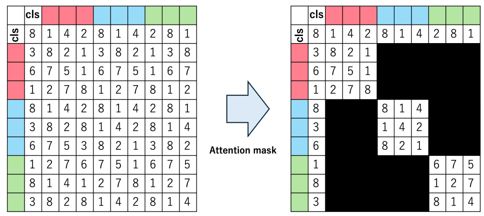

Fig. 2. Structure of the Attention Mask [8].

The learnable threshold τ is constrained via a loss function to approach this target z value:

$$\mathscr{L}_{th} = \lambda \left| \tau - (\mu + z_r \sigma) \right|, \tag{15}$$

where λ is a weighting factor. This design allows the model to automatically learn an optimal pruning threshold that achieves the desired removal rate for each sample (Fig. 4).

Additionally, passing the scores through a sigmoid function normalizes them to $[0,1]$, stabilizing gradients. The gradients are computed directly with respect to τ, enabling learnable frame selection.

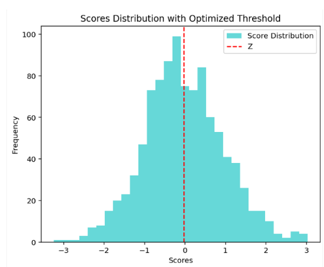

Fig. 3. Score Normal Distribution and Target Variable z.

3.2.2 Top-K Segment Method for Fixed Threshold Frame Selection

In the proposed method, we adopt a "Top-K segment" approach for fixed threshold frame selection, where the video is divided into multiple temporal segments, and the top K frames within each segment are selected. This ensures that important frames are evenly distributed across the video, allowing temporal information to be considered in frame pruning.

The procedure is as follows.

1. Fixed Threshold Setting. The mean of all frame importance scores s_f is used as the threshold:

$$\theta = \frac{1}{F} \sum_{f=1}^{F} s_f, \tag{16}$$

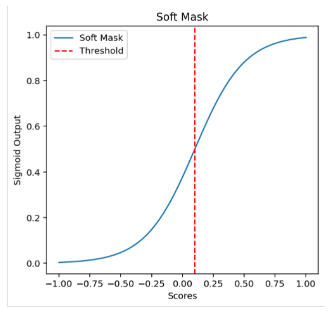

Fig. 4. Schematic Diagram of the Inverse Function of the Cumulative Distribution Function (CDF) of the Normal Distribution and Threshold.

where F is the total number of frames. This threshold serves as a reference for normalizing the scores into a soft mask.
2. Soft Mask Computation. Each frame score s_f is converted into a soft mask using the sigmoid function:

$$\text{soft\_mask}_f = \sigma\big(s \cdot (s_f - \theta)\big), \tag{17}$$

where σ is the sigmoid function and s is the slope parameter. This normalizes scores relative to the threshold, emphasizing the relative importance of frames.
3. Top-K Segment Selection. The video is divided into n temporal segments, and the top K frames within each segment (based on the soft mask) are selected. This guarantees a uniform distribution of important frames across the video and allows subtle actions and temporal variations to be considered in the pruning process.

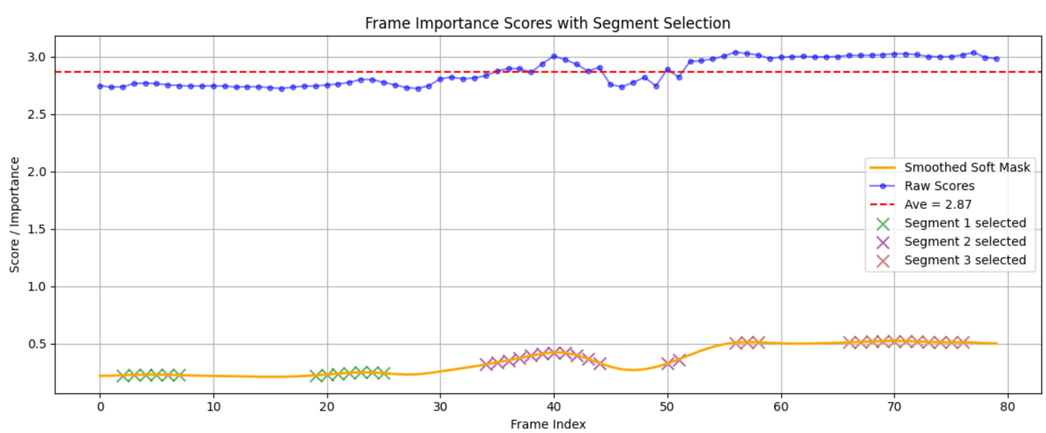

Fig. 5. Overview of Top-K Segment Selection.

The Top-K segment method is novel in that it not only selects frames exceeding the threshold but also ensures temporal uniformity. Furthermore, the soft mask emphasizes relative frame importance, enabling efficient selection of the most informative frames.

4 Experiments

In this experiment, we train the ViViT-Tiny model using the proposed method described in the previous section and evaluate its performance by comparing recognition accuracy, FLOPs, and the number of parameters. This experiment validates the effectiveness of multi-modal analysis and importance-based frame pruning in action recognition using ViViT.

4.1 Dataset

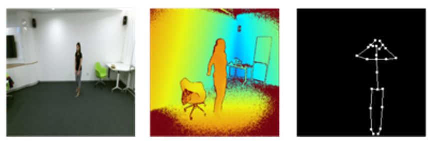

Fig. 6. Example of Dataset.

The dataset used in this experiment consists of NTU RGB+D [9] and NTU RGB+D 120 [10], which summarize behaviors that occur in human daily life. This dataset is composed of a large amount of action data including individuals of various ages. Specifically, it contains 114,480 video samples across 120 different actions. In this study, we will use RGB video, and by generating depth and skeleton videos through the estimation model, we will utilize three types of videos. We will select 10 types of actions (standing, sitting, falling, headache, coughing, punching, kicking, lying down, running, and picking up objects) and split 900 data samples per class into training and testing sets at a ratio of 7:3.

4.2 Experimental Settings

We summarize the hyperparameters and implementation details used in this study. All models were trained with input frames of size 224×224, and a batch size of 1 was employed due to memory constraints associated with video data. The Adam optimizer was used with a learning rate of 1×10^{-5}, and training was conducted for 100 epochs using the cross-entropy loss function. For the learnable threshold method, the pruning rate was set to 50%. For the Top-K method, the top 50% of frames within each segment were selected to maintain consistency. All experiments were conducted under the same environment to ensure fair comparison across different settings.

4.3 Evaluation Metrics

In this experiment, we use accuracy, number of parameters, and computational complexity (FLOPs) as metrics to evaluate the performance of each model in action recognition. Additionally, we visualize the remaining frames after pruning to assess which method performs more effective frame selection.

5 Results

In this study, to verify the effectiveness of the proposed Multi-modal Action Density Scoring (MADS) as a method for reducing computational cost, a series of experiments were conducted using two different frame selection strategies. We compared the performance of ViViT models of various sizes when applying frame pruning based on the two proposed frame selection methods, evaluating them in terms of accuracy, number of parameters, and computational complexity (FLOPs).

5.1 Quantitative Results

Table 1 summarizes the quantitative results of the proposed methods applied to different ViViT model sizes. We compare three configurations: the baseline ViViT without pruning, the Learnable Threshold method, and the proposed Top-K Segment Method. The pruning rate was set between 20–36%, and evaluation was performed using accuracy, number of parameters, and FLOPs.

From the results, it can be observed that the Top-K Segment Method achieves a significant reduction in FLOPs (up to 36%) while maintaining almost the same accuracy as the baseline model. In particular, the accuracy drop remained within only 0.5% across all model sizes, demonstrating that this method preserves critical temporal information during pruning.

Table 1: Comparison of ViViT models with different frame selection methods. FLOPs are adjusted according to the specified change percentages. Accuracy shows relative drop (%).

Model Size	Method	Pruning Rate (%)	Accuracy (%)	Params (M)	FLOPs (G)	Change in FLOPs (%)
ViViT-Tiny	Baseline	–	77.02	7.26	86.42	–
	Learnable Threshold	20	74.75 (−2.95%)	7.26	**70.61**	(−18.3%)
	Top-K Segment	24	77.01 (−0.01%)	7.26	**65.68**	(−24.0%)
ViViT-Small	Baseline	–	83.75	23.56	209.21	–
	Learnable Threshold	25	81.10 (−3.16%)	23.56	**162.14**	(−22.5%)
	Top-K Segment	31	83.54 (−0.25%)	23.56	**144.35**	(−31.0%)
ViViT-Base	Baseline	–	90.69	89.21	520.67	–
	Learnable Threshold	30	88.98 (−1.89%)	89.21	**374.36**	(−28.1%)
	Top-K Segment	36	90.22 (−0.52%)	89.21	**333.23**	(−36.0%)

On the other hand, the Learnable Threshold method exhibits a larger degradation in recognition accuracy (up to −3.2%), despite similar levels of computational reduction. This indicates that the learned threshold may not effectively capture the temporal distribution of important frames, leading to biased selection and reduced representational quality.

Overall, these quantitative results confirm that the Top-K Segment approach achieves a better balance between accuracy and efficiency compared to the Learnable Threshold method.

5.2 Qualitative Results

To further analyze the behavior of each frame selection method, we visualize the remaining frames after pruning and their corresponding frame importance heatmaps, as shown in Fig. 7.

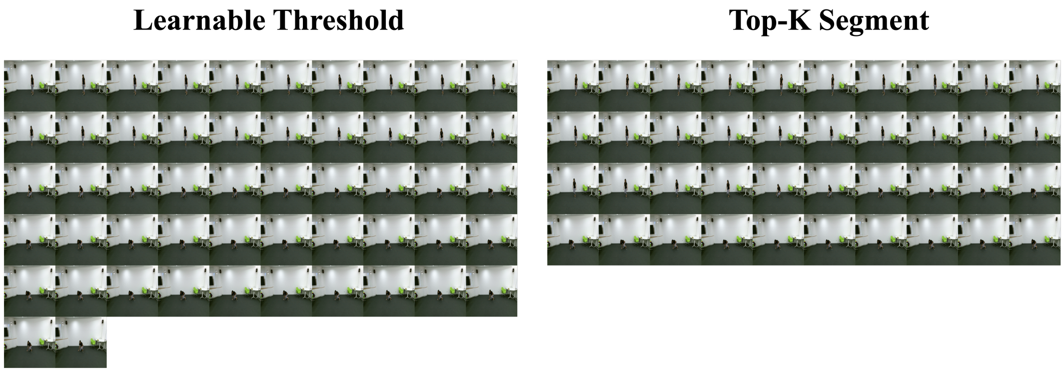

Fig. 7. Comparison Diagram Using Two Methods.

In the case of the Learnable Threshold method, the selected frames were biased toward the beginning and end of the video sequence, with most intermediate frames being discarded. This indicates that the learned threshold tended to overfit to temporal boundary cues, leading to a loss of meaningful motion continuity and degraded recognition performance.

In contrast, the Top-K Segment method retained frames more evenly across the entire video sequence. The corresponding heatmaps show that important regions were distributed throughout the video, covering diverse motion phases and interaction events without strong temporal bias. This balanced selection demonstrates that Top-K Segment effectively captures temporally salient information while preserving representational diversity, contributing to both high recognition accuracy and improved computational efficiency (Fig. 8).

Fig. 8. Heatmap Comparison Diagram.

Overall, these qualitative observations confirm that the proposed Top-K Segment method provides a more stable and interpretable pruning behavior compared to the Learnable Threshold approach.

6 Discussion

The experimental results demonstrate that the proposed Multi-modal Action Density Scoring (MADS) framework effectively reduces computational cost while maintaining recognition accuracy across ViViT models of different sizes. In particular, the Top-K Segment Method achieved significant FLOPs reduction (up to 36%) with almost no accuracy degradation (less than 0.5%), whereas the Learnable Threshold Method exhibited larger accuracy drops of up to 3%.

This difference can be attributed to the underlying frame selection mechanism. The Learnable Threshold method tends to select only temporally concentrated regions—often the beginning or the end of a video—resulting in a loss of temporal diversity. In contrast, the Top-K Segment method divides a video into temporal segments and selects the most informative frames from each segment. This allows for a more uniform temporal coverage, ensuring that key motion cues are retained throughout the sequence.

Qualitative analysis, supported by visualizations of the retained frames and importance heatmaps, further confirms this behavior. The Learnable Threshold method retains frames mainly concentrated in specific regions, while the Top-K Segment method distributes selected frames more evenly, capturing motion transitions and maintaining contextual continuity. These findings highlight that considering temporal segmentation is crucial for robust and efficient frame pruning in video recognition.

7 Conclusion

In this study, we presented a novel temporal frame selection framework for ViViT-based action recognition leveraging Multi-modal Action Density Scoring (MADS). To address computational bottlenecks in high-dimensional video data, we introduced and evaluated two mechanisms: the Learnable Threshold and the Top-K Segment Method. Our investigation focused on balancing model efficiency and predictive performance in real-world applications.

Experimental results demonstrated that the Top-K Segment Method significantly outperforms the Learnable Threshold in both stability and efficiency. Specifically, the Top-K method achieved a substantial 36% reduction in FLOPs while keeping accuracy degradation within 0.5% of the baseline. In contrast, the Learnable Threshold suffered a more pronounced accuracy drop of up to 3%, likely due to its tendency to overlook intermediate temporal features.

Qualitative analysis via importance heatmaps validated that the Top-K Segment approach ensures a more uniform and representative sampling of frames. By preserving the continuity of motion and interaction events, this method maintains the representational integrity of Video Vision Transformers, confirming that structured temporal segmentation is essential for robust frame pruning.

Future work will extend the MADS framework toward adaptive multi-modal fusion, dynamically adjusting to modalities like audio or optical flow. We also plan to explore token-level merging to further optimize models for resource-constrained edge devices. Ultimately, this research provides a scalable foundation for efficient, high-performance large-scale video understanding.

References

[1] Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C. ViViT: A Video Vision Transformer. In: International Conference on Computer Vision (ICCV); 2021. .

[2] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.

[3] Han S, Pool J, Tran J, Dally WJ. Learning both weights and connections for efficient neural network. In: Advances in Neural Information Processing Systems (NeurIPS); 2015. .

[4] Kim J, Park J, Kim K. Temporal Feature Alignment for Video Action Recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021. .

[5] Xu H, Xie L, Liu W. Temporal Action Detection with Similarity-based Frame Pruning. In: International Conference on Multimedia and Expo (ICME); 2019. .

[6] Zhou B, Andonian A, Oliva A, Torralba A. Temporal Video Segmentation for Human-centric Action Recognition. In: European Conference on Computer Vision (ECCV); 2018. .

[7] Baltrusaitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018;41(2):423-43.

[8] Higashi R Takashi Ishibashi, Meng L. Action Recognition Using MEViViT:Modality Embedding ViViT. In: IAI2025 The 7th International Conference on Industrial Artificial Intelligence; 2025. .

[9] Shahroudy A, Liu J, Ng TT, Wang G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 1010-9.

[10] Liu J, Shahroudy A, Perez M, Wang G, Duan LY, Kot AC. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence. 2019;42(10):2684-701.