

On Modulating Motion-Aware Visual-Language Representation for Few-Shot Action Recognition

Pengfei Fang^{ID}, Qiang Xu^{ID}, Zixuan Lin^{ID}, and Hui Xue^{ID}, *Member, IEEE*

Abstract—This paper focuses on few-shot action recognition (FSAR), where the machine is required to understand human actions, with each only seeing a few video samples. Even with only a few explorations, the most cutting-edge methods employ the action textual features, pre-trained by a visual-language model (VLM), as a cue to optimize video prototypes. However, the action textual features used in these methods are generated from a static prompt, causing the network to overlook rich motion cues within videos. To tackle this issue, we propose a novel framework, namely, motion-aware visual-language representation modulation network (MoveNet). The proposed MoveNet utilizes dynamic motion cues within videos to integrate motion-aware textual and visual feature representations, as a way to modulate the video prototypes. In doing so, a long short motion aggregation module (LSMAM) is first proposed to capture diverse motion cues. Having the motion cues at hand, a motion-conditional prompting module (MCPM) utilizes the motion cues as conditions to boost the semantic associations between textual features and action classes. One further develops a motion-guided visual refinement module (MVRM) that adopts motion cues as guidance in enhancing local frame features. The proposed components compensate for each other and contribute to significant performance gains over the FASR task. Thorough experiments on five standard benchmarks demonstrate the effectiveness of the proposed method, considerably outperforming current state-of-the-art methods.

Index Terms—Few-shot action recognition, visual-language model, motion modeling.

I. INTRODUCTION

DUE to the increasing demand for video understanding in real-world applications, the field of video action recognition has made significant progress [1], [2], [3], [4] in recent years. However, most of these studies rely on a large amount of annotated data to learn video representations, which is time-consuming and labor-intensive. Therefore,

Received 23 October 2024; revised 26 January 2025 and 5 March 2025; accepted 27 March 2025. Date of publication 2 April 2025; date of current version 8 September 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62306070 and Grant 62476056, in part by the Social Development Science and Technology Project of Jiangsu Province under Grant BE2022811, in part by Southeast University Start-Up Grant for New Faculty under Grant 4009002309, and in part by the Big Data Computing Center of Southeast University. This article was recommended by Associate Editor R. Hong. (Pengfei Fang and Qiang Xu contributed equally to this work.) (Corresponding author: Hui Xue.)

Pengfei Fang, Qiang Xu, and Hui Xue are with the School of Computer Science and Engineering and the Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, Ministry of Education, Southeast University, Nanjing 211189, China (e-mail: fangpengfei@seu.edu.cn; 220232307@seu.edu.cn; hxue@seu.edu.cn).

Zixuan Lin is with the School of Astronautics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: 2213210011@stu.xjtu.edu.cn).

Digital Object Identifier 10.1109/TCSVT.2025.3557009

few-shot action recognition (FSAR) is a promising direction to alleviate dependence on massive labeled data and has received widespread attention [5], [6], [7], [8], [9]. FSAR aims to classify unlabeled query videos into one of the action classes in the support set using only a few samples.

Current mainstream FSAR approaches adopt metric learning paradigm [10] and episode training to optimize the model, which maps input videos into a common feature space and measures the alignment distances between query and support prototypes in each specific task. Previous methods [11], [12], [13], [14], which typically utilize single-modal pre-trained backbones, focus on designing novel prototype generation modules or effective alignment metrics. Despite notable improvements, these methods struggle to leverage single visual modality without incorporating multi-modal knowledge, limiting their further performance gains. Fortunately, contrastive language-image pre-training, *a.k.a.*, CLIP, adopts the contrastive learning scheme to acquire transferable multi-modal knowledge, bridging the semantic gap between the two modalities and generating complementary feature representations. Since CLIP has achieved significant performance on various downstream tasks [15], [16], [17], [18], a few works attempt to apply the CLIP model to the FSAR task [19], [20], [21]. One of the most representative works is CLIP-FSAR [19], which employs action textual features to modulate reliable and robust support prototypes, as shown in Figure 1(a). This work is remarkable as it is the first to enhance visual representations using textual features in the realm of FSAR.

Despite their substantial performance gains, these studies, such as CLIP-FSAR [19], fail to explore the critical role of motion information (*i.e.*, motion cues) in video understanding [23], [24], [25]. Motion cues encapsulate diverse contextual information inherent in videos due to their temporal nature, which reflect the evolution of action changes rather than static individual frames. Although previous methods [13], [26], [27] attempt to extract motion features (*i.e.*, temporal-difference features) by computing the differences between adjacent frames, they primarily focus on low-rate action changes, which refers to the gradual transitions in actions between adjacent frames. In real-world scenarios, actions evolve smoothly over time, leading to minimal differences between adjacent frames, as illustrated in Figure 2. This is particularly evident in cases where the action is slow or the frame rate is high. Moreover, CLIP-FSAR [19] relies on static prompt forms, *e.g.*, “A photo of [CLS]” to generate textual features. Using this static form, the network is inadequately equipped to capture the temporal dynamics within videos, especially with limited samples under

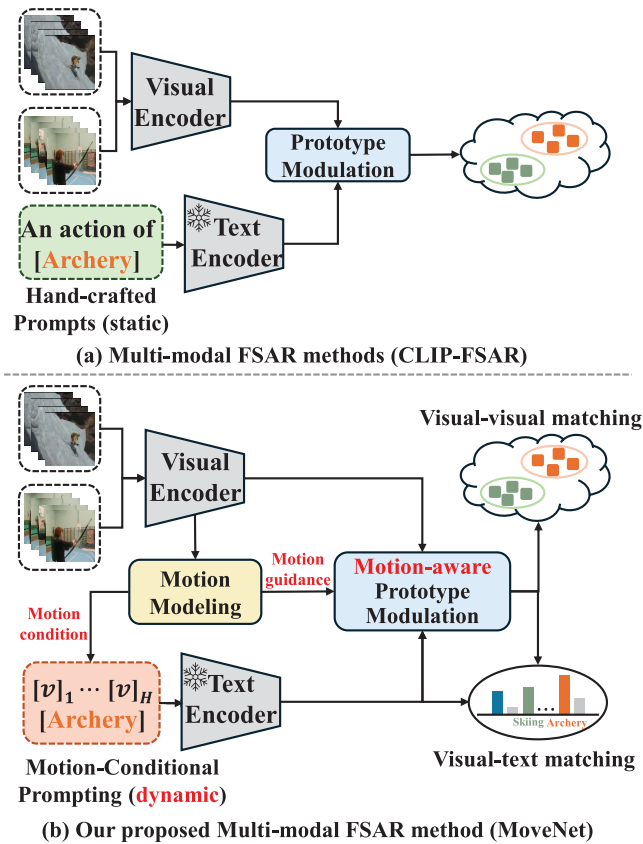


Fig. 1. Motivation of our work. Comparison with previous FSAR methods. (a) Existing multi-modal FSAR methods (e.g., CLIP-FSAR) primarily utilize textual features for visual prototype modulation through a straightforward mechanism; (b) Our proposed MoveNet introduces a motion-conditioned prompting module that dynamically generates textual features conditioned on motion cues, enabling effective modulation of motion-aware multi-modal representations.

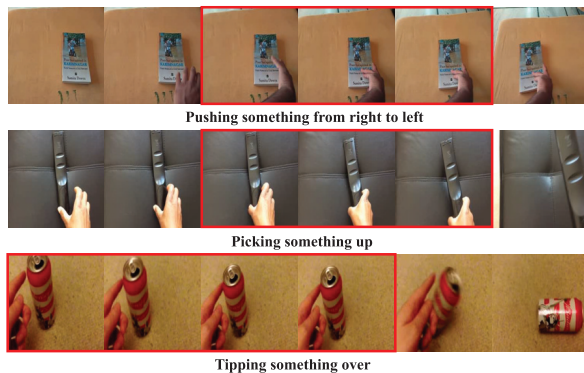


Fig. 2. Illustration of samples from SSv2-Full [22]. We use red bounding boxes to highlight adjacent frames that exhibit high visual similarity.

few-shot settings. While the CLIP model learns high-quality transferable representations by applying contrastive learning on image-text pairs, it struggles with video data. One of the key challenges is its inability to model temporal relationships effectively. CLIP processes individual frame inputs in isolation, which fails to capture the temporal dynamics within videos. Additionally, pre-training a language-video model is resource-expensive due to large-scale video-text data and

extensive training resources. An alternative approach is to endow the image-based CLIP model to understand dynamic visual information in videos. Since several works [28], [29], [30] have explored the integration of visual information into prompt design, we believe a similar idea could be applied to generate dynamic prompts based on motion cues within videos.

Inspired by these insights, we propose a novel motion-aware visual-language representation modulation network (MoveNet) to utilize dynamic motion cues within videos to comprehensively modulate motion-aware textual and visual feature representations, as illustrated in Figure 1(b). Specifically, to extract meaningful motion cues to represent entire action changes, we introduce a long short motion aggregation module (LSMAM) to capture and aggregate both long-range and short-range motion cues, modeling dynamic contextual information within videos. With motion cues available, a novel motion-conditional prompting module (MCPM) is incorporated to leverage motion cues as conditions and enable the text encoder to produce motion-related textual features, thereby boosting their semantic associations with corresponding action classes. The importance of different local frame features varies in overall action classification. Due to this reason, we propose a motion-guided visual refinement module (MVRM) to further employ the motion cues as guidance in enhancing local frame features since they contain diverse contextual information about the whole action. In this way, our proposed MoveNet effectively utilizes dynamic motion cues within videos to modulate textual and visual feature representations. These are both motion-aware and complementary to each other, thereby achieving robust and accurate action classification. In summary, our major contributions are as follows:

- We propose a novel motion-aware visual-language representation modulation network (MoveNet) for few-shot action recognition, which effectively modulates both textual and visual feature representations by utilizing dynamic motion cues.
- We design a long short motion aggregation module (LSMAM) to capture dynamic motion cues within videos comprehensively. To fully utilize the motion cues, a motion-conditional prompting module (MCPM), and a motion-guided visual refinement module (MVRM) are introduced. The former employs motion cues as conditions to produce motion-related textual features, and the latter adopts motion cues as guidance to refine visual frame features.
- Extensive experiments have been conducted across five standard benchmarks to demonstrate that the proposed MoveNet is effective and achieves state-of-the-art performance.

II. RELATED WORK

In this section, we briefly review several research topics relevant to our paper, including few-shot action recognition, motion learning, and visual-language pretraining.

A. Few-Shot Action Recognition

Few-shot action recognition (FSAR) aims to classify unlabeled query videos into one of the action classes in the support

set. Unlike conventional action recognition [23], [31], which focuses on recognizing actions from a pre-defined set of action classes with large amounts of labeled training data for each class. The FSAR task focuses on recognizing novel actions with only a few labeled examples per class. Most existing FSAR methods, such as OTAM [12] and TRX [11], adhere to the metric learning paradigm. These approaches seek to learn a common feature space where alignment distances between query and support videos can be computed for video action classification using a pre-defined temporal alignment metric.

Early methods [32], [33], [34] usually condense frame-level features into a single feature representation for the entire video, which overlooks the dynamic temporal information. Several recent works focus on designing novel temporal alignment metrics. For instance, OTAM [12] adopts a variant of dynamic time warping (DTW) algorithm to utilize the temporal ordering information for aligning query and support videos. TRX [11] matches videos by matching each sub-sequence of the query video with all sub-sequences in the support set, while HyRSM [14] introduces a bidirectional Mean Hausdorff Metric to facilitate flexible temporal alignment in the matching process. Other approaches aim to enhance frame features for FSAR [5], [6], [7], [26], [35]. MTFAN [26] focuses on temporal segment alignment, and STRM [35] utilizes a local patch-level enrichment module and a global frame-level enrichment module to learn high-order feature representations. Despite their great success, these methods utilize single-modal pre-trained backbones without incorporating multi-modal knowledge, which limits their further performance gains. In our paper, we propose a novel multi-modal framework to modulate textual and visual feature representations that are complementary, contributing to more accurate few-shot action classification.

B. Motion Learning

Motion information refers to the temporal information in videos, which captures the dynamic action changes over time. Motion cues have been widely recognized as crucial for action recognition [36] as they provide diverse contextual information that helps distinguish different actions, especially when static appearance features alone are insufficient. Several action recognition methods [37], [38], [39], [40], [41] utilize 3D CNNs to simultaneously extract both spatial and temporal features, but this approach is resource-intensive due to the large number of parameters involved. Another line of approaches utilizes two-stream networks to capture motion features from optical flow. However, the required additional networks and the high computational cost of optical flow still limit the efficiency of these methods. To address these challenges, recent approaches [27], [42] have developed temporal difference modules that extract motion features directly from raw frame inputs, achieving promising results and serving as an alternative to optical flow for capturing motion cues. In this paper, we also focus on extracting motion cues through a specialized module. Unlike previous methods, we introduce a long short motion aggregation module that effectively captures and aggregates both long-range and short-range motion cues into a comprehensive motion vector, which is then used to modulate motion-aware textual and visual feature representations.

C. Vision-Language Pretraining

Vision-Language Pretraining (VLP) has emerged as a prominent research focus for its remarkable ability to bridge visual content with textual descriptions. One of the most influential works is CLIP [43], which maps images and natural language descriptions into a common feature space for contrastive learning. Due to its exceptional zero-shot transferability, several studies have explored adapting the CLIP model for diverse image-related downstream tasks, such as vision-and-language navigation [44], [45], pathology detection [46] and image classification [28], [47]. For instance, [44], [45] integrates commonsense knowledge and instructions to enhance visual representations for the navigation task. Reference [46] aligns disease descriptions curated by experts with medical images to improve pathology detection. However, pre-training such a language-video model like CLIP is resource-expensive since it requires large-scale video-text data pairs and extensive training resources. Due to this, transferring language-image pre-trained models to the video domain has received widespread attention due to its training efficiency. For instance, ActionCLIP transforms the recognition task into a video-text matching problem and proposes a novel “pre-train, prompt, fine-tune” paradigm. X-CLIP [48] introduces a cross-frame module into the CLIP image encoder for temporal modeling. MoTED [31] leverages advanced large language models (LLMs) to generate discriminative motion-related descriptions. MUPPET [29] further introduces a Multi-Modality Prompt Meta-Learning method to generate multi-modal prompts to address the challenge of few-shot temporal action localization (FS-TAD). Inspired by these successes, several works [19], [20] have attempted to adapt the CLIP model for the FSAR task. While existing multi-modal FSAR methods, such as CLIP-FSAR [43] focus on employing textual features to modulate visual prototypes, our work explores the potential of integrating visual information into the text branch.

III. METHOD

In this section, we first introduce the problem formulation of few-shot action recognition. Then we present the overall architecture of the proposed MoveNet. Finally, we describe each key component of our model.

A. Problem Formulation

In the FSAR task, the objective is to classify unlabeled query videos into one of the N action classes in the support set, where each action class has only K samples available. This setting is commonly referred to as an N -way K -shot problem. In a typical few-shot learning setting, the entire dataset is divided into two disjoint subsets, a base dataset \mathcal{D}_{train} for training and a novel dataset \mathcal{D}_{test} for testing, which is different from traditional classification tasks [23], [27], [31] where training and testing share the same set of classes. Following the common practice of episode training paradigm [11], [12], [35], [49], the episode is randomly selected from \mathcal{D}_{train} in the training phase to optimize the model. In each episode, a support set $\mathcal{S} = \{s_1, s_2, \dots, s_{N \times K}\}$ is constructed, consisting of $N \times K$ samples drawn from N different action classes with

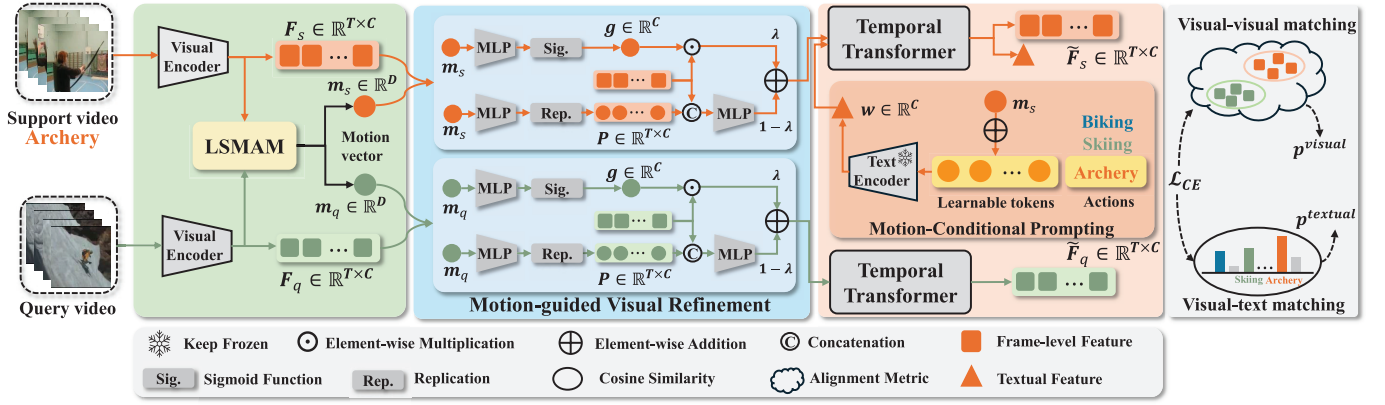


Fig. 3. The Architecture of our proposed MoveNet under the 1-way 1-shot setting. Given the support and query videos, we employ a visual encoder to encode frame features. Then we apply a long short motion aggregation module (LSMAM) to capture and aggregate dynamic motion cues within videos. With motion cues available, a motion-conditional prompting module is utilized to leverage motion cues as conditions to allow textual features to learn action-related information. Further, a motion-guided visual refinement module (MVRM), followed by a temporal transformer module is employed to enhance frame features under the guidance of motion cues and to perform interactions between frame features. Finally, a pre-defined temporal alignment metric is imposed on the query feature and support prototypes for classification.

K video samples per class where $s_i \in \mathbb{R}^{T \times 3 \times H \times W}$. The task aims to classify the query videos in the query set \mathcal{Q} based on the information provided by the samples in the support set.

B. Architecture Overview

The overall architecture of our proposed MoveNet is illustrated in Figure 3. For simplicity, we describe our framework using an N -way 1-shot task, where the query set \mathcal{Q} contains a single query video q . Following previous works [11], [12], [50], we implement a sparse frame sampling strategy on input videos to reduce redundant information and computational burden. In this setup, the support set can be denoted as $\mathcal{S} = \{s_1, s_2, \dots, s_N\}$. We employ the visual encoder of CLIP to extract the query feature F_q and support features $F_S = \{F_{s_1}, F_{s_2}, \dots, F_{s_N}\}$, where $F_i = \{f_i^1, f_i^2, \dots, f_i^T\}, f_i^j \in \mathbb{R}^C$. Motion information, which represents the differences between two successive frames, captures the dynamic action changes in videos and is crucial for accurate action recognition [36]. To capture motion, we propose a long short motion aggregation module (LSMAM) which extracts the differences between consecutive frame features and aggregates them into a comprehensive motion vector that contains dynamic action changes throughout the entire video. With motion cues available, a motion-conditional prompting module (MCPM) is designed to leverage the motion cues as conditions to integrate dynamic motion information into textual features, thereby enhancing their semantic associations with corresponding action classes. Additionally, we propose a motion-guided visual refinement module (MVRM) to refine local frame features under the guidance of motion cues. A temporal transformer module in MoLo [13] is then employed to facilitate interactions among different frame features and integrate multi-modal textual semantic information into visual features of the support set. Finally, we compute the distances between the query feature and the support prototypes using a pre-defined temporal alignment metric for classification. Moreover, we incorporate the cosine similarity between the query features and the textual features corresponding to the action classes in the support set as part of the final classification.

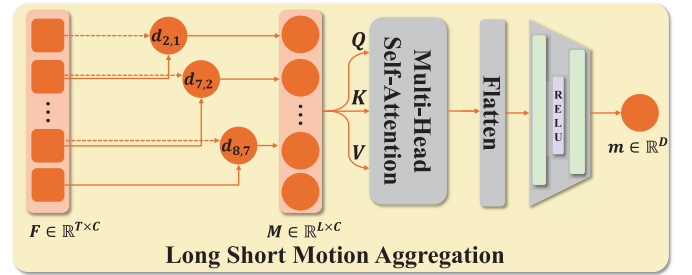


Fig. 4. Illustration of long short motion aggregation module (LSMAM). LSMAM first captures long-range and short-range motion cues within videos and subsequently aggregates them into a comprehensive motion vector.

C. Long Short Motion Aggregation

Previous methods [13], [26], [27] primarily focus on low-rate action changes, the gradual transitions in actions between adjacent frames, which are insufficient to fully describe the dynamics of the action, leading to sub-optimal performance. To address this limitation, we propose a long short motion aggregation module (LSMAM), which captures both long-range and short-range motion cues from frame feature representations and aggregates them into a comprehensive motion vector, as illustrated in Figure 4.

Specifically, we compute the forward frame feature differences from the input video features $F = \{f^1, f^2, \dots, f^T\}$ (subscript is omitted for simplicity) using a hyper-parameter s ($1 \leq s \leq T - 1$), which represents the maximum temporal interval between two frame features. Then a set of frame feature differences M , can be represented as:

$$M = \{d_{j,i} = f^j - f^i \mid 1 \leq i < j \leq T, j - i \leq s\} \in \mathbb{R}^{L \times C}, \quad (1)$$

where $L = \sum_{i=1}^s (T - i)$ represents the number of frame feature differences (*i.e.*, motion cues). Then we feed the frame feature differences M into a multi-head self-attention (MHSA) mechanism to enhance the interactions between long-range and short-range motion cues:

$$\hat{M} = M + \text{MHSA}(M). \quad (2)$$

The resulting motion features $\hat{\mathbf{M}}$, are first flattened into a vector $\hat{\mathbf{m}} \in \mathbb{R}^{LC}$. To aggregate the motion cues and reduce the dimension of $\hat{\mathbf{m}}$, a linear projection is applied, followed by a non-linear activation function. Subsequently, another linear projection is used to adjust the motion vector, given by:

$$\mathbf{m} = \mathbf{W}_2^T(\text{ReLU}(\mathbf{W}_1^T(\hat{\mathbf{m}}))), \quad (3)$$

where $\mathbf{W}_1 \in \mathbb{R}^{LC \times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C \times D}$. Here, D represents the dimension of the final comprehensive motion vector \mathbf{m} , which effectively encapsulates diverse motion cues throughout the entire video.

D. Motion-Conditional Prompting

To enable the CLIP model to learn dynamic motion information from video data, a prompt learner [51] is used to generate dynamic prompts conditioned by motion cues. Concretely, the prompt learner defines the prompt with the following form:

$$\mathcal{P} = [\mathbf{v}]_1[\mathbf{v}]_2 \cdots [\mathbf{v}]_H[\text{CLASS}][.], \quad (4)$$

where each $[\mathbf{v}]_i (i \in \{1, 2, \dots, H\})$ is a learnable token vector with the same dimension D as the word embeddings, and H is a hyper-parameter representing the number of learnable tokens. To allow these token vectors to learn action-related semantic information, the comprehensive motion vector \mathbf{m} is employed as a conditional token and is combined with token vectors $[\mathbf{v}]_i$ through element-wise addition:

$$[\mathbf{v}]_i = [\mathbf{v}]_i \oplus \mathbf{m}, i \in \{1, 2, \dots, H\}, \quad (5)$$

where \oplus denotes element-wise addition. The resulting dynamic prompts are then fed into the text encoder of CLIP, denoted as \mathcal{T} , to generate textual features corresponding to each action class in the support set. To preserve the powerful multi-modal transferable knowledge within the text encoder, \mathcal{T} remains frozen during the training phase.

In a standard N -way 1-shot task, the support set consists of N action classes, with each containing only one support video. However, in an N -way K -shot task, each class contains K support videos, which share the same action labels but exhibit different motion information. To this end, we generate specific textual features for each video in the support set. For each video j belonging to action class i , the corresponding textual feature is denoted as $w_{i,j}$, where $1 \leq i \leq N$ and $1 \leq j \leq K$. Thus, the set of all textual features is:

$$\mathcal{W} = \{w_{i,j} \mid 1 \leq i \leq N, 1 \leq j \leq K\}. \quad (6)$$

By generating these textual features with the incorporation of motion cues, we significantly boost their semantic associations with corresponding action classes, thereby allowing for a more detailed understanding of actions.

E. Motion-Guided Visual Refinement

Different local frame features encapsulate distinct action changes at various temporal positions throughout the entire video. To enable local frame features to perceive global motion dynamics, we propose a motion-guided visual refinement module (MVRM). This module comprises two components: motion-guided channel reweighting (MCR) and motion-guided

feature integration (MFI). Together, they refine and enhance individual frame features under the guidance of dynamic motion cues, providing a comprehensive understanding of the entire action.

1) *Motion-Guided Channel Reweighting*: The MCR aims to emphasize the feature channels that are most relevant to the action. Different channels of frame features can vary significantly for different actions. To address this, we assign weights to the feature channels based on the motion cues throughout the entire video. Given the input video features $\mathbf{F} = \{\mathbf{f}^1, \mathbf{f}^2, \dots, \mathbf{f}^T\} \in \mathbb{R}^{T \times C}$, and the pre-extracted comprehensive motion vector $\mathbf{m} \in \mathbb{R}^D$, the motion vector is first passed through an MLP block to adjust its dimension:

$$\mathbf{m}' = \text{MLP}(\mathbf{m}) \in \mathbb{R}^C, \quad (7)$$

where the MLP block consists of two linear layers with a ReLU activation function. A Sigmoid activation function is then applied to obtain a channel-wise weight vector \mathbf{g} :

$$\mathbf{g} = \text{Sigmoid}(\mathbf{m}') \in \mathbb{R}^C, \quad (8)$$

Finally, the weight vector \mathbf{g} is multiplied element-wise with the original frame features \mathbf{F} to emphasize channels relevant to the motion cues:

$$\mathbf{F}_{MCR} = \mathbf{F} \odot \mathbf{g} + \mathbf{F}. \quad (9)$$

2) *Motion-Guided Feature Integration*: The objective of MFI is to refine and enhance individual frame features by incorporating global motion information. The underlying idea is that the motion vector captured by LSMAM contains diverse motion information throughout the entire video, closely related to the action class. To be specific, given the input video features $\mathbf{F} = \{\mathbf{f}^1, \mathbf{f}^2, \dots, \mathbf{f}^T\} \in \mathbb{R}^{T \times C}$ and the corresponding comprehensive motion vector $\mathbf{m} \in \mathbb{R}^D$, the motion vector is processed through an MLP block to adjust its dimension:

$$\mathbf{m}'' = \text{MLP}(\mathbf{m}) \in \mathbb{R}^C. \quad (10)$$

The adjusted motion vector \mathbf{m}'' is then replicated T times to match the number of frame features:

$$\mathbf{P} = \text{Replicate}(\mathbf{m}'', T) \in \mathbb{R}^{T \times C}. \quad (11)$$

The replicated motion features \mathbf{P} are concatenated with the input features \mathbf{F} :

$$\mathbf{F}_{concat} = \text{Concat}(\mathbf{F}, \mathbf{P}) \in \mathbb{R}^{T \times 2C}, \quad (12)$$

where Concat represents the concatenation operation along the channel dimension. A second MLP block is then used to integrate the motion cues into local frame features with a residual connection to retain the original features:

$$\mathbf{F}_{MFI} = \text{MLP}(\mathbf{F}_{concat}) + \mathbf{F}. \quad (13)$$

The final output features of the whole motion-guided visual refinement module is the weighted summation of the two refined features \mathbf{F}_{MCR} and \mathbf{F}_{MFI} with a hyper-parameter $\lambda \in [0, 1]$, given by:

$$\tilde{\mathbf{F}} = \lambda \mathbf{F}_{MCR} + (1 - \lambda) \mathbf{F}_{MFI}. \quad (14)$$

F. Prediction Scheme

In the FSAR task, existing methods primarily follow the metric learning paradigm, which measures the alignment distances between query and support prototypes through a pre-defined metric for classification. Inspired by the successful approach in CLIP-FSAR [19], a temporal transformer module is first used to leverage action textual features to enhance support visual features and further obtain reliable and robust prototypes.

$$[\mathbf{w}_{i,1}, \tilde{\mathbf{F}}_{s_i}] = \text{Tformer}([\mathbf{w}_{i,1}, \tilde{\mathbf{F}}_{s_i}]), i \in \{1, 2, \dots, N\}, \quad (15)$$

where $\mathbf{w}_{i,1} \in \mathbb{R}^C$ denotes the corresponding textual feature to $\tilde{\mathbf{F}}_{s_i}$ and Tformer represents the temporal transformer module in CLIP-FSAR [19]. Here, $[\cdot, \cdot]$ denotes concatenation along the temporal dimension. Since the query label is unknown during the training phase, we simply feed the query video feature $\tilde{\mathbf{F}}_q$ into this module:

$$\tilde{\mathbf{F}}_q = \text{Tformer}(\tilde{\mathbf{F}}_q). \quad (16)$$

The final support video features and query video feature are denoted as $\tilde{\mathbf{F}}_S = \{\tilde{\mathbf{F}}_{s_1}, \tilde{\mathbf{F}}_{s_2}, \dots, \tilde{\mathbf{F}}_{s_N}\}$, and $\tilde{\mathbf{F}}_q$, respectively. We implement the temporal alignment metric on these features:

$$d_{q,s_i} = \mathcal{M}(\tilde{\mathbf{F}}_q, \tilde{\mathbf{F}}_{s_i}), \quad (17)$$

where \mathcal{M} represents the pre-defined temporal alignment metric. In our proposed method, the metric \mathcal{M} is TRX [11] by default. Based on the distance d_{q,s_i} , we can obtain the probability distribution over support classes for query video q :

$$P_{(y=i|q)}^{visual} = \frac{\exp(d_{q,s_i})}{\sum_{j=1}^N \exp(d_{q,s_j})}. \quad (18)$$

The CLIP model leverages powerful multi-modal knowledge to align visual features with their corresponding textual features. Building on this, our proposed motion-conditional prompting module and motion-guided visual refinement module contribute to modulating motion-aware textual and visual feature representations informed by dynamic motion cues, making them complementary to each other. Consequently, query videos can be classified not only with support visual prototypes but also with the corresponding textual features from the support set.

Given the textual feature outputs $\mathcal{W} = \{\mathbf{w}_{i,j} \mid 1 \leq i \leq N, 1 \leq j \leq K\}$ where $K = 1$ in our discussion, we calculate the average of textual features for support videos that belong to the same action class to obtain the representative textual feature $\bar{\mathbf{w}}_i$ for each action class i :

$$\bar{\mathbf{w}}_i = \frac{1}{K} \sum_{j=1}^K \mathbf{w}_{i,j}, 1 \leq i \leq N. \quad (19)$$

The final set of representative textual features for all action classes in the support set is denoted as:

$$\bar{\mathcal{W}} = \{\bar{\mathbf{w}}_i \mid 1 \leq i \leq N\}. \quad (20)$$

Given the final support features $\tilde{\mathbf{F}}_S = \{\tilde{\mathbf{F}}_{s_1}, \tilde{\mathbf{F}}_{s_2}, \dots, \tilde{\mathbf{F}}_{s_N}\}$ and query feature $\tilde{\mathbf{F}}_q$, where $\tilde{\mathbf{F}}_v = \{\tilde{\mathbf{f}}_v^1, \tilde{\mathbf{f}}_v^2, \dots, \tilde{\mathbf{f}}_v^T\}$, $v \in \{s_1, s_2, \dots, s_N, q\}$, we calculate the cosine similarity between each frame feature and textual features rather than using global

average-pooling (GAP) as each frame contains distinct action details. The probability distribution is given by:

$$P_{(y=i|v)}^{textual} = \frac{\exp\left(\left(\frac{1}{T} \sum_{j=1}^T \text{sim}(\tilde{\mathbf{f}}_v^j, \bar{\mathbf{w}}_i)\right) / \tau\right)}{\sum_{k=1}^N \exp\left(\left(\frac{1}{T} \sum_{j=1}^T \text{sim}(\tilde{\mathbf{f}}_v^j, \bar{\mathbf{w}}_k)\right) / \tau\right)}, \quad (21)$$

where $v \in \{s_1, s_2, \dots, s_N, q\}$. The final prediction of query video q can be represented as:

$$P_{(y=i|q)} = P_{(y=i|q)}^{visual} + \alpha \cdot P_{(y=i|q)}^{textual}, \quad (22)$$

where $\alpha \in [0, 1]$ is a hyper-parameter. Following previous works [11], [12], [13], we employ a standard cross-entropy loss \mathcal{L}_{CE} to optimize the model parameters.

IV. EXPERIMENTS

In this section, we first introduce five standard FSAR benchmarks (*i.e.*, SSv2-Full, SSv2-Small [22], Kinetics [59], UCF101 [60], and HMDB51 [61]). Then we describe the implementation details of our MoveNet. After that, we compare our method with current state-of-the-art FSAR methods on five standard benchmarks. Finally, we perform thorough experiments to demonstrate the effectiveness of the proposed MoveNet.

A. Datasets

We evaluate the effectiveness of the proposed MoveNet on five standard FSAR benchmarks, including SSv2-Full [22], SSv2-Small, Kinetics [59], UCF101 [60] and HMDB51 [61]. For SSv2-Full and SSv2-Small, we follow the data splits presented in OTAM [12] and CMN [53], where 100 classes are divided into 64/12/24 non-overlapping action classes for training, validation, and testing. SSv2-Full includes all samples from the original dataset, while SSv2-Small consists of 100 samples per class. Regarding Kinetics [59], we adopt the few-shot benchmark provided by CMN [53], which includes 100 randomly selected classes from the original Kinetics dataset. The UCF101 [60] dataset contains 101 action classes and we follow the few-shot splits in ARN [62]. For HMDB51 [61], the common practice is to split 31 classes for training and 10 classes for testing.

It is noted that a significant portion of video samples in the SSv2-Full and SSv2-Small datasets emphasizes temporal dynamics of actions (*e.g.*, “opening something” and “dropping something into something”). In contrast, the Kinetics, UCF101, and HMDB51 datasets are more focused on recognizing specific action classes, such as “jumping,” and “playing drums.”

B. Implementation Details

1) *Network Architecture*: Since most of existing FSAR methods [11], [13], [35], [49], [58] employ the ResNet-50 [63] pre-trained on ImageNet [64] as the visual encoder, we adopt the ResNet-50 in the CLIP (*i.e.*, CLIP-ResNet-50) as the visual encoder in our study. To preserve the powerful transferable knowledge from the visual-language model, we keep the text encoder of CLIP frozen throughout the training phase.

TABLE I

COMPARISON TO RECENT FSAR METHODS ON SSV2-FULL, SSV2-SMALL, KINETICS, UCF101 AND HMDB51. THE EXPERIMENT IS CONDUCTED UNDER THE 5-WAY 1-SHOT AND 5-WAY 5-SHOT SETTINGS. "INET-RN50" DENOTES RESNET-50 PRE-TRAINED ON IMAGENET AND "CLIP-RN50" DENOTES RESNET-50 IN THE CLIP. "-" MEANS THE RESULT IS NOT AVAILABLE IN PUBLISHED WORKS. "*" STANDS FOR THE RESULTS OF OUR IMPLEMENTATION. THE BEST RESULTS ARE GIVEN IN BOLD

Method	Reference	Pre-training	SSv2-Full		SSv2-Small		Kinetics		UCF101		HMDB51	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MatchingNet [10]	NeurIPS'16	INet-RN50	-	-	31.3	45.5	53.3	74.6	-	-	-	-
MAML [52]	ICML'17	INet-RN50	-	-	30.9	41.9	54.2	75.3	-	-	-	-
CMN [53]	ECCV'18	INet-RN50	-	-	36.2	48.8	57.3	76.0	-	-	-	-
OTAM [12]	CVPR'20	INet-RN50	42.8	52.3	36.4	48.0	73.0	85.8	79.9	88.9	54.5	68.0
ITANet [54]	IJCAI'21	INet-RN50	49.2	62.3	39.8	53.7	73.6	84.3	-	-	-	-
TRX [11]	CVPR'21	INet-RN50	42.0	64.6	36.0	56.7	63.6	85.9	78.2	96.1	53.1	75.6
TA ² N [55]	AAAI'22	INet-RN50	47.6	61.0	-	-	72.8	85.8	81.9	95.1	59.7	73.9
MTFAN [26]	CVPR'22	INet-RN50	45.7	60.4	-	-	74.6	87.4	84.8	95.1	59.0	74.6
STRM [35]	CVPR'22	INet-RN50	43.1	68.1	37.1	55.3	62.9	86.7	80.5	96.9	52.3	77.3
HyRSM [14]	CVPR'22	INet-RN50	54.3	69.0	40.6	56.1	73.7	86.1	83.9	94.7	60.3	76.0
Huang <i>et al.</i> [56]	ECCV'22	INet-RN50	49.3	66.7	38.9	61.6	73.3	86.4	71.4	91.0	60.1	77.0
Nguyen <i>et al.</i> [57]	ECCV'22	INet-RN50	43.8	61.1	-	-	74.3	87.4	84.9	95.9	59.6	76.9
MoLo [13]	CVPR'23	INet-RN50	55.0	69.6	41.9	56.2	74.0	85.6	86.0	95.5	60.8	77.4
SloshNet [49]	AAAI'23	INet-RN50	46.5	68.3	-	-	70.4	87.0	86.0	97.1	59.4	77.5
GgHM [58]	ICCV'23	INet-RN50	54.5	69.2	-	-	74.9	87.4	85.2	96.3	61.2	76.9
TRX (Baseline) [11]	CVPR'21	CLIP-RN50	49.7*	68.2*	39.4	56.8	76.1*	87.4*	86.9*	96.2*	62.4	80.3*
CLIP-Freeze [43]	ICML'21	CLIP-RN50	28.5	37.8	26.8	36.3	68.2	85.3	84.6	94.5	51.4	71.0
CLIP-FSAR [19]	IJCV'23	CLIP-RN50	58.1	62.8	52.0	55.8	87.6	91.9	91.3	97.0	69.2	80.3
MVP-shot [20]	ArXiv'24	CLIP-RN50	-	-	51.2	57.0	90.0	93.2	92.2	97.6	72.5	82.5
Ours (MoveNet)	-	CLIP-RN50	61.9	72.7	53.4	65.0	90.6	93.5	94.8	97.9	75.7	86.1

2) *Training and Inference*: Similar to previous methods [11], [12], [50], we sparsely and uniformly sample 8 frames to encode video representations, *i.e.*, $T = 8$. Each frame is first resized to 256×256 . During the training phase, we then apply standard data augmentation strategies, including 224×224 random cropping, horizontal flipping, and normalization. For the testing phase, we simply use a center cropping. To evaluate the few-shot performance on each benchmark, we follow previous works [11], [12], [19] and report the average accuracy over 10,000 episodes randomly sampled from the test set under the 5-way 1-shot and 5-way 5-shot settings.

C. Comparison With State-of-the-Art Methods

To thoroughly verify the effectiveness of our proposed method, we first compare the performance of our MoveNet over current state-of-the-art FSAR methods on five standard benchmarks. The comparison results are presented in Table I.

1) *Comparison With TRX Baseline*: We compare the proposed MoveNet with the TRX baseline method and replace the visual encoder in TRX with the pre-trained CLIP-ResNet-50 for a fair comparison. As a result, our method consistently achieves significant improvements across all five benchmarks, particularly under the 5-way 1-shot setting. Notably, we observe substantial gains of 12.2%, 14.0%, 14.5%, 7.9%, and 13.3% on SSv2-Full, SSv2-Small [22], Kinetics [59], UCF101 [60], and HMDB51 [61] datasets, respectively.

2) *Comparison With Multi-Modal Approaches*: By leveraging the multi-modal knowledge of CLIP, several multi-modal approaches, *e.g.*, CLIP-FSAR [19] and MVP-Shot [20], have outperformed previous state-of-the-art single-modal methods [13], [14], [49]. Building on this foundation, our method also achieves significant performance gains. Compared with CLIP-FSAR [19], the first notable work to

apply the CLIP model to the FSAR task, our MoveNet achieves superior results across all five benchmarks under both 1-shot and 5-shot settings. For instance, under the 5-way 1-shot setting, our approach yields improvements of 3.8%, 3.0%, 3.5%, and 6.5% gains on the SSv2-Full, Kinetics, UCF101, and HMDB51 datasets, respectively. Furthermore, our MoveNet consistently outperforms MVP-Shot, particularly on the SSv2-Small and HMDB51 datasets. Under the 5-way 1-shot setting, our method achieves performance gains of 2.2% and 3.2% on SSv2-Small and HMDB51, respectively. Additionally, our method yields gains of 8.0% and 3.6% on these two datasets under the 5-way 5-shot setting.

D. Ablation Study

We conduct extensive ablation experiments to validate the ability of our proposed MoveNet and to analyze the contribution of each individual component. Unless otherwise specified, we use the pre-trained CLIP-ResNet-50 as the default visual encoder to ensure a fair comparison.

1) *Analysis of Each Component*: To validate the contributions of each component in our MoveNet, we conduct experiments under the 5-way 5-shot and 5-way 1-shot settings on the HMDB51 [61] and UCF101 [60] datasets. As shown in Table II, we examine the roles and contributions of two key components: motion-conditional prompting module (MCPM) and motion-guided visual refinement module (MVRM). The MCPM alone yields improvements of 8.5% and 2.9% on HMDB51 under the 1-shot and 5-shot settings, respectively, demonstrating its ability to utilize dynamic motion cues as conditions to generate textual features that are closely associated with corresponding action classes. Additionally, the MVRM brings performance increases of 3.8% and 0.9% over the baseline on UCF101, highlighting the importance

TABLE II

ABLATION STUDY ON EACH COMPONENT IN OUR MOVENET. WE REPORT THE 5-WAY 1-SHOT AND 5-WAY 5-SHOT RESULTS ON HMDB51 [61] AND UCF101 [60] DATASETS. THE BEST RESULTS ARE GIVEN IN BOLD

MCPM	MVRM	HMDB51		UCF101	
		1-shot	5-shot	1-shot	5-shot
✗	✗	62.4	80.3	86.9	96.2
✓	✗	70.9	83.2	92.6	96.8
✗	✓	65.6	83.3	90.7	97.1
✓	✓	75.7	86.1	94.8	97.8

TABLE III

ABLATION STUDY ON DIFFERENT MOTION MODELING METHODS. WE REPORT THE 5-WAY 1-SHOT AND 5-WAY 5-SHOT RESULTS ON HMDB51 [61] AND UCF101 [60]. THE BEST RESULTS ARE GIVEN IN BOLD

Motion modeling	HMDB51		UCF101	
	1-shot	5-shot	1-shot	5-shot
TAP	69.4	81.8	93.2	96.9
Temporal Difference + TAP	72.1	82.9	93.6	97.0
LSMAM w/o MHSA	69.7	81.2	94.0	97.1
LSMAM	75.7	86.1	94.8	97.9

of motion cues in video action recognition. Notably, the performance gains are more significant under the 1-shot setting, indicating that textual semantic information becomes more effective when visual information is limited, and our method efficiently leverages motion cues throughout videos to enhance the model’s generalization performance. Furthermore, the combination of these two components results in substantial gains and achieves state-of-the-art performance, suggesting that our proposed MCPM and MVRM are highly complementary and together contribute significantly to the overall effectiveness.

2) *Analysis of Motion Modeling*: In this paper, we introduce a novel motion modeling method, namely, long short motion aggregation. To verify its motion modeling capability, we conduct experiments to analyze the impact of different motion modeling methods. The results are shown in Table III. “TAP” denotes we simply take the average of frame features as the motion vector. “Temporal Difference + TAP” means that we compute the differences between adjacent frames and take the average of these temporal difference features as the motion vector. “LSMAM w/o MHSA” denotes we replace the attention mechanism in LSMAM with average pooling. Experimental results demonstrate that averaging frame features or frame their differences as motion cues fails to effectively describe global action dynamics within videos, leading to sub-optimal performance. Furthermore, by applying an attention mechanism to enhance interactions between motion cues, our approach shows significant improvements, particularly on the HMDB51 [61] dataset, where we achieve gains of 6.0% and 4.9% under 5-way 1-shot and 5-way 5-shot settings, respectively.

3) *Analysis of Motion-Guided Visual Refinement*: The aim of the motion-guided visual refinement module (MVRM) is to comprehensively refine and enhance local frame features

TABLE IV

ABLATION STUDY ON TWO COMPONENTS IN MVRM. WE REPORT THE 5-WAY 1-SHOT AND 5-WAY 5-SHOT RESULTS ON HMDB51 [61] AND UCF101 [60] DATASETS. THE BEST RESULTS ARE GIVEN IN BOLD

MVRM		HMDB51		UCF101	
MCR	MFI	1-shot	5-shot	1-shot	5-shot
Multi-Head Attention		72.3	84.4	92.7	97.1
✗	✗	70.9	83.2	92.6	96.8
✓	✗	73.3	85.6	93.4	97.1
✗	✓	73.6	85.2	93.9	97.3
✓	✓	75.7	86.1	94.8	97.8

TABLE V

ABLATION STUDY ON DIFFERENT INITIALIZATIONS OF LEARNABLE TOKENS IN MCPM. “w/ m ” AND “w/o m ” DENOTE WHETHER MOTION CUES ARE UTILIZED. THE BEST RESULTS ARE GIVEN IN BOLD

Initialization $H = 4$	HMDB51		UCF101	
	1-shot	5-shot	1-shot	5-shot
A video of action [CLS] w/o m	73.7	83.5	93.1	97.3
A video of action [CLS] w/ m	75.2	83.8	93.9	97.6
$[v]_1[v]_2 \dots [v]_H$ [CLS] w/o m	70.0	81.1	92.5	97.0
$[v]_1[v]_2 \dots [v]_H$ [CLS] w/ m	75.7	86.1	94.8	97.8

under the guidance of motion cues, which consists of two components: motion-guided channel reweighting (MCR) and motion-guided feature integration (MFI). We conduct experiments on the HMDB51 [61] and UCF101 [60] datasets to validate the effectiveness of these two components. As shown in Table IV, applying MCR and MFI separately results in performance improvements. When applied individually, MCR and MFI achieve gains of 2.4% and 2.7%, respectively, under the 5-way 1-shot setting on HMDB51. Furthermore, jointly utilizing MCR and MFI yields even greater performance improvements, with considerable gains of 4.8% on HMDB51 and 1.7% on UCF101 under the 1-shot setting. We further replace the MCR and MFI with a Multi-Head Attention operation, followed by a feedforward layer to facilitate interactions between motion features and frame features. These results demonstrate that our proposed MVRM effectively leverages motion information to enhance individual frame features to perceive global motion dynamics and outperforms the attention-based approach across all datasets and settings, underscoring its effectiveness.

4) *Initialization of Learnable Tokens*: To further explore the impact of different initializations of the learnable token vectors in MCPM, we conduct experiments under the 5-way 1-shot and 5-way 5-shot settings on the HMDB51 [61] and UCF101 [60] datasets. We compare the performance of hand-crafted prompts (“A video of action [CLS]”) with randomly initialized prompts, and examine whether to use motion cues as conditions. As shown in Table V, without motion cues, hand-crafted prompts outperform randomly initialized prompts, e.g., 73.7% vs. 70.0% on HMDB51 under the 1-shot setting, and 97.3% vs. 97.0% on UCF101 under the 5-shot setting. However, incorporating motion cues as conditions for dynamic textual feature generation leads to significant performance improvements of 2.3% on HMDB51 under the 5-shot setting, and 0.4% on UCF101 under the 1-shot setting. This suggests

TABLE VI

ABLATION STUDY ON DIFFERENT NUMBERS OF LEARNABLE TOKENS H IN MCPM. THE BEST RESULTS ARE GIVEN IN BOLD

Number of learnable tokens H	HMDB51		UCF101	
	1-shot	5-shot	1-shot	5-shot
2	73.4	84.5	93.2	97.6
4	75.7	86.1	94.8	97.8
8	72.3	85.0	93.5	97.9
16	75.4	85.3	92.8	97.7

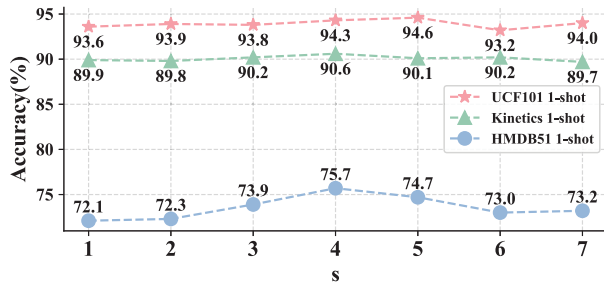


Fig. 5. The impact of the hyper-parameter s , which represents the maximum temporal interval between two frame features on Kinetics [59], UCF101 [60] and HMDB51 [61] datasets under the 5-way 1-shot setting.

that in FSAR tasks, randomly initialized prompts, lacking structured constraints, are more susceptible to over-fitting due to limited data and the motion conditions are effective in dynamic textual feature generation.

5) *Number of Learnable Tokens*: We compare the results of different numbers of learnable tokens in MCPM under the 5-way 1-shot and 5-way 5-shot settings on the HMDB51 [61] and UCF101 [60] datasets. The results are displayed in Table VI, which demonstrate that selecting an appropriate number of learnable tokens H can further enhance the effectiveness of our proposed MCPM. We achieve state-of-the-art performance with 86.1% on HMDB51 when $H = 4$ and 97.9% on UCF101 when $H = 8$ under the 5-way 5-shot setting. These results indicate that a small value of H may lead to insufficient information learned by MCPM, while a large value may result in redundant information, both of which negatively impact the final classification accuracy.

6) *Impact of Temporal Interval s in LSMAM*: We conduct a sensitivity analysis for hyper-parameter s in the proposed LSMAM to investigate the contributions of short-range and long-range motion cues. A small value of s indicates that the model primarily focuses on short-range motion cues, while a large value of s enables the model to capture feature differences over a broader temporal range. As illustrated in Figure 5, when the value of s is small, the model can only capture short-range motion cues, which are insufficient to fully describe the dynamics of the action, leading to sub-optimal performance, especially on HMDB51 where the accuracy is only 72.1% when $s = 1$. Conversely, when s is large, the performance also declines. This can be attributed to the model capturing redundant motion information due to the excessive temporal range, which degrades performance. Therefore, selecting an appropriate value of s (e.g., 4 and 5) allows LSMAM to effectively capture both short-range and long-range motion cues, leading to a more comprehensive understanding of action changes throughout the video.

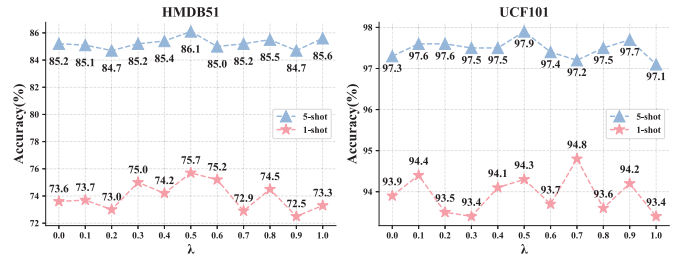


Fig. 6. The impact of the varying weighting factor λ in MVRM on UCF101 [60] and HMDB51 [61] under 5-way 1-shot and 5-way 5-shot settings.

TABLE VII

GENERALIZATION EXPERIMENTS WITH DIFFERENT TEMPORAL ALIGNMENT METRICS ON KINETICS [59] AND SSV2-SMALL [22] UNDER 5-WAY 1-SHOT AND 5-WAY 5-SHOT SETTINGS

Temporal alignment metric	Kinetics		SSv2-small	
	1-shot	5-shot	1-shot	5-shot
OTAM [12]	76.4	88.9	39.9	51.9
CLIP-FSAR (OTAM) [19]	87.6	91.9	52.0	55.8
Ours (OTAM)	89.9	91.3	52.4	54.5
BiMHM [14]	77.1	88.9	40.6	52.2
CLIP-FSAR (BiMHM) [19]	87.7	92.1	52.3	55.6
Ours (BiMHM)	90.3	91.5	53.7	54.9
TRX [11]	76.1	87.4	39.4	56.8
CLIP-FSAR (TRX) [19]	86.7	92.2	51.5	57.1
Ours (TRX)	90.6	93.5	53.4	65.0

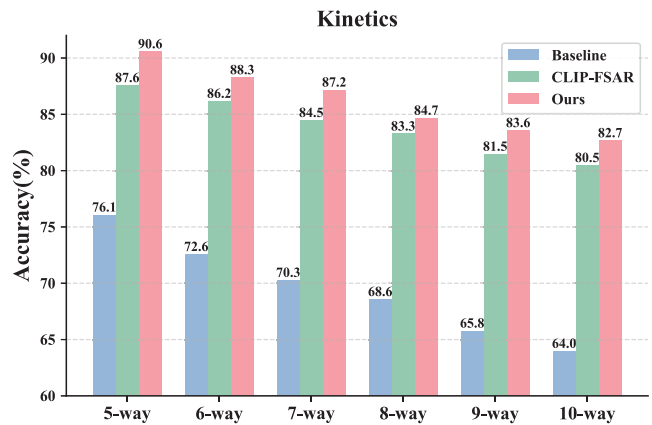


Fig. 7. N -way 1-shot results of the proposed MoveNet, CLIP-FSAR [19] and TRX [11] baseline on Kinetics [59] dataset with N ranging from 5 to 10. The visual encoder is pre-trained CLIP-ResNet-50.

7) *Impact of Weighting Factor λ in MVRM*: We conduct experiments on the HMDB51 [61] and UCF101 [60] datasets to validate the contributions of these two components, MCR and MFI by adjusting the weighting factor λ in MVRM. As shown in Figure 6, the model exhibits robustness to small variations in λ within the range of 0.0 to 0.4, where performance remains relatively stable. Additionally, for lower values of λ , where MFI (motion-guided feature integration) dominates, the model achieves high performance, suggesting that MFI leads to better generalization and robustness. As λ increases beyond 0.6, the influence of MCR (motion-guided channel reweighting) increases, leading to a gradual

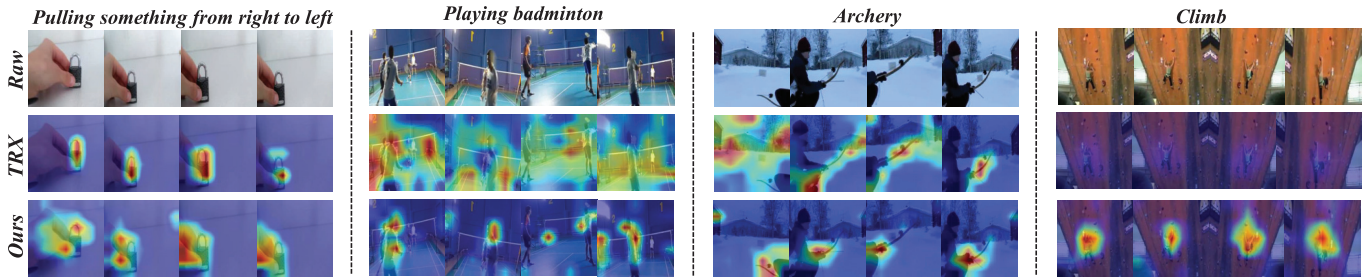


Fig. 8. Attention visualization of our MoveNet with Grad-CAM [66] under the 5-way 1-shot setting. Top three rows correspond to classes from SSv2-Full and SSv2-Small [22], and the bottom three are selected from Kinetics [59], UCF101 [60], and HMDB51 [61], respectively. Corresponding to the original video frames, the attention maps of TRX are compared to the attention maps of our MoveNet.

TABLE VIII

ABLATION STUDY ON DIFFERENT BACKBONES. WE REPLACE THE PRE-TRAINED CLIP-RESNET-50 WITH THE PRE-TRAINED CLIP-ViT-B/16 AND X-CLIP-ViT-B/16 [48], AND CONDUCT QUANTITATIVE EXPERIMENTS ON SSv2-FULL, SSv2-SMALL [22], KINETICS [59], UCF101 [60] AND HMDB51 [61]. “-” MEANS THE RESULT IS NOT AVAILABLE IN PUBLISHED WORKS. THE BEST RESULTS ARE GIVEN IN BOLD

Method	Pre-training	SSv2-Full		SSv2-Small		Kinetics		UCF101		HMDB51	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
CLIP-FSAR [19]	CLIP-RN50	58.1	62.8	52.0	55.8	87.6	91.9	91.3	97.0	69.2	80.3
MVP-shot [20]	CLIP-RN50	-	-	51.2	57.0	90.0	93.2	92.2	97.6	72.5	82.5
Ours (MoveNet)	CLIP-RN50	61.9	72.7	53.4	65.0	90.6	93.5	94.8	97.9	75.7	86.1
CLIP-FSAR [19]	CLIP-ViT-B	61.9	72.1	54.5	61.8	89.7	95.0	96.6	99.0	75.8	87.7
MVP-shot [20]	CLIP-ViT-B	-	-	55.4	62.0	91.0	95.1	96.8	99.0	77.0	88.1
Ours (MoveNet)	CLIP-ViT-B	62.7	73.6	56.3	64.4	91.2	95.3	96.2	99.0	77.9	88.4
Ours (MoveNet)	X-CLIP-ViT-B	63.7	74.5	56.6	66.9	93.9	97.4	96.8	99.1	74.8	86.3

decline in performance. This indicates that excessive emphasis on MCR may compromise the model’s stability.

8) *Generalization on Different Alignment Metrics*: We conduct experiments with three different temporal alignment metrics: OTAM [12], BiMHM [14], and TRX [11] to demonstrate the generalization ability of the proposed MoveNet. OTAM [12] explicitly estimates an ordered temporal alignment path to match the frames of two videos. BiMHM [14] formulates the distance measure between two videos as a set matching problem and designs a Bi-directional Mean Hausdorff Metric to align frames. TRX [11] matches actions by matching plentiful ordered tuples of different sub-sequences. As shown in Table VII, our method achieves impressive performance, particularly with the TRX metric [11]. Compared with baseline methods, our method demonstrates significant improvements, indicating its effectiveness and versatility across various temporal alignment metrics.

9) *Performance of N-Way 1-Shot Classification*: To further validate the robustness of our MoveNet, we conduct experiments on the Kinetics dataset [59] under N -way 1-shot settings, where N varies from 5 to 10. A larger value of N represents more action classes within a task, increasing the difficulty of action classification. As shown in Figure 7, the performance under the 10-way setting decreases by 7.9% compared to that under the 5-way setting. Despite this, the performance of our MoveNet still outperforms other comparison methods across various settings, demonstrating that by leveraging additional textual semantic information and diverse motion cues, our approach effectively handles more challenging few-shot scenarios.

10) *Influence of Different Backbones*: Previous FSAR methods [11], [13], [35] utilize a ImageNet [64] pre-trained ResNet-50 as the basic visual encoder. To evaluate the performance of different backbones, we replace the pre-trained CLIP-ResNet-50 with CLIP-ViT/B-16 as the visual encoder and conduct experiments across all five benchmarks (*i.e.*, SSv2-Full, SSv2-Small [22], Kinetics [59], UCF101 [60] and HMDB51 [61]). Furthermore, transferring language-image pre-trained models to the video domain [31], [48], [65] has achieved significant attention due to its striking performance and training efficiency. Hence, we also adopt a language-video pre-trained model, X-CLIP [48] as the visual encoder, which has been trained fully-supervised on Kinetics-400 [59]. The results, presented in Table VIII, show that using CLIP-ViT and X-CLIP-ViT as the visual encoder significantly outperforms the ResNet-50-based approach, demonstrating that a strong pre-trained backbone leads to improved generalization.

E. Visualization Analysis

1) *Attention Visualization*: Figure 8 illustrates the attention visualization our MoveNet under the 5-way 1-shot setting. Corresponding to the input video frames, the attention maps of the TRX baseline, are compared to the attention maps generated by our MoveNet. We select a temporal-related action class, “Pulling something from left to right” from the SSv2-Small [22] dataset. Additionally, three spatial-related actions, “Playing badminton,” “Archery,” and “Climb” are selected from Kinetics [59], UCF101 [60], and HMDB51 [61], respectively. As shown in Figure 8, the attention maps generated by our MoveNet effectively highlight action changes and

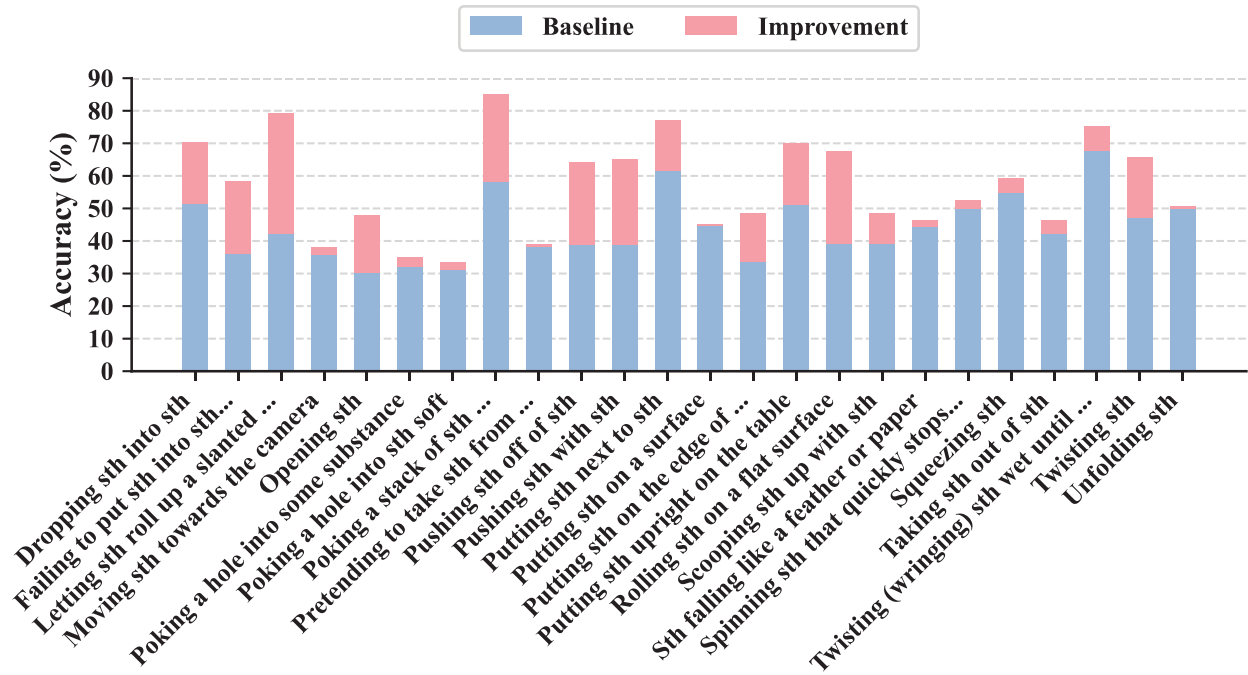


Fig. 9. Visualization of performance gains of different action classes in the SSv2-Full [22] dataset under the 5-way 1-shot setting.

action-related objects. For instance, in the action of “Pulling something from right to left,” the attention map focuses on the movement of the hand holding the lock rather than the lock itself. Furthermore, in the action of “Playing badminton,” the fourth attention map successfully captures the serve motion, underscoring the critical motion cues within the entire action.

2) *Class Improvement Visualization*: To further demonstrate the motion modeling capabilities of our proposed model, we conduct experiments on the temporal-sensitive dataset SSv2-Full [22]. Specifically, we performed quantitative experiments to investigate the performance improvements across different action classes compared to the TRX [11] baseline. The results are visualized in Figure 9. It can be observed that our proposed model achieves significant performance gains across a wide range of action classes in the SSv2-Full test set. Notably, our model achieves over 20% improvement on highly motion-sensitive classes, such as “Dropping something into something,” “Pushing something off of something” and “Putting something next to something.” These consistent improvements on motion-sensitive actions underscore the effectiveness of our model in capturing and modeling motion cues within videos.

3) *Similarity Visualization*: To further qualitatively analyze the impact of each component of our proposed MoveNet, we visualize the similarities between query and support videos in an episode from the Kinetics [59] dataset under 5-way 1-shot setting. As shown in Figure 10, integrating MCRM and MVRM separately into TRX baseline enhances the discrimination of video features, leading to more accurate classifications. Furthermore, the proposed MoveNet, which combines MCPM and MVRM, achieves even more precise and robust classification decisions, particularly for the first and the fourth query samples (q_1, q_4). These results further demonstrate that the proposed two modules, MCPM and MVRM are effective and complementary.

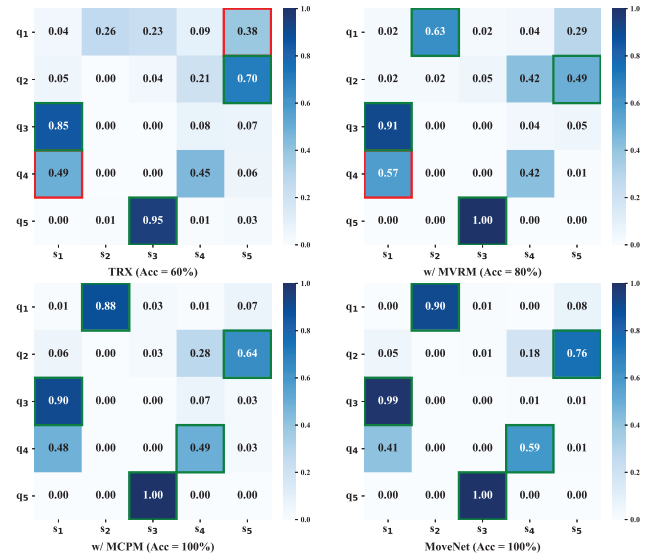


Fig. 10. Similarity visualization between query samples (q_n) and support videos (s_n) in an episode from Kinetics [59]. “w/ MVRM” and “w/ MCPM” denote that we integrate our proposed MVRM and MCPM into TRX baseline separately. A higher score indicates a greater degree of similarity. The green box indicates correct prediction and the red box indicates incorrect prediction.

4) *Semantic Similarity Visualization*: We conduct experiments to verify whether the motion-conditioned textual features retain their original semantic integrity. Specifically, we calculate cosine similarities between the textual features generated by the proposed MoveNet and the original CLIP model [43]. As depicted in Figure 11, the diagonal values of the similarity matrix represent the semantic similarities between two sets of textual features. The results show that most diagonal values exceed 0.8 and even surpass 0.9, indicating that the motion-conditioned textual features generated

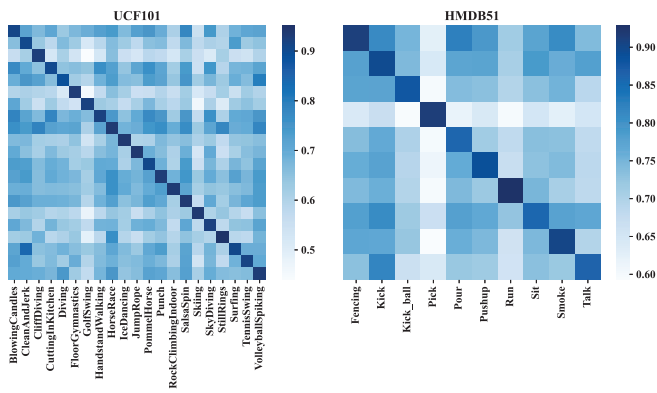


Fig. 11. Visualization of semantic similarity matrix between the textual features obtained by the proposed MoveNet and the original CLIP [43] on UCF101 [60] and HMDB51 [61].

by our MoveNet preserve their correct semantic information and maintain semantic integrity.

V. CONCLUSION

In this work, we propose a novel motion-aware visual-language representation modulation network (MoveNet) to address the challenge of few-shot action recognition. The proposed MoveNet effectively leverages diverse motion cues to generate textual features that are closely aligned with specific action classes. It also introduces a motion-guided visual refinement module to enhance individual frame features, thereby modulating motion-aware textual and visual feature representations to achieve accurate action classification. We believe that our MoveNet offers a new perspective on integrating visual dynamics into the textual branch, which will inspire further research. In future work, we will continue to explore the application of visual dynamics within Vision Transformers (ViTs).

REFERENCES

- J. Weng, C. Weng, J. Yuan, and Z. Liu, "Discriminative spatio-temporal pattern discovery for 3D action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 1077–1089, Apr. 2019.
- H. Wang and L. Wang, "Cross-agent action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2908–2919, Oct. 2018.
- X. Jiang, K. Xu, and T. Sun, "Action recognition scheme based on skeleton representation with DS-LSTM network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 2129–2140, Jul. 2020.
- J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2019, pp. 7083–7093.
- X. Wang et al., "Task-aware dual-representation network for few-shot action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 10, pp. 5932–5946, Oct. 2023.
- X. Wang, Y. Lu, W. Yu, Y. Pang, and H. Wang, "Few-shot action recognition via multi-view representation learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 9, pp. 8522–8535, Sep. 2024.
- M. Lu, S. Yang, X. Lu, and J. Liu, "Cross-modal contrastive pre-training for few-shot skeleton action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 10, pp. 9798–9807, Oct. 2024.
- S. Liu, M. Jiang, and J. Kong, "Multidimensional prototype refactor enhanced network for few-shot action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6955–6966, Oct. 2022.
- Z. Yang, G. An, Z. Zheng, S. Cao, and Q. Ruan, "GBC: Guided alignment and adaptive boosting CLIP bridging vision and language for robust action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 9, pp. 8172–8187, Sep. 2024.
- O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, vol. 29, Dec. 2016, pp. 3637–3645.
- T. Perrett, A. Masullo, T. Burghardt, M. Mirmehdi, and D. Damen, "Temporal-relational CrossTransformers for few-shot action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 475–484.
- K. Cao, J. Ji, Z. Cao, C.-Y. Chang, and J. C. Niebles, "Few-shot video classification via temporal alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10618–10627.
- X. Wang et al., "MoLo: Motion-augmented long-short contrastive learning for few-shot action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18011–18021.
- X. Wang et al., "Hybrid relation guided set matching for few-shot action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19916–19925.
- R. Zhang et al., "Tip-adapter: Training-free CLIP-adapter for better vision-language modeling," 2021, *arXiv:2111.03930*.
- P. Gao et al., "CLIP-adapter: Better vision-language models with feature adapters," 2021, *arXiv:2110.04544*.
- Y. Rao et al., "DenseCLIP: Language-guided dense prediction with context-aware prompting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18061–18070.
- H. Shi, M. Hayat, Y. Wu, and J. Cai, "ProposalCLIP: Unsupervised open-category object proposal generation via exploiting CLIP cues," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9601–9610.
- X. Wang et al., "CLIP-guided prototype modulating for few-shot action recognition," *Int. J. Comput. Vis.*, vol. 132, no. 6, pp. 1899–1912, Jun. 2024.
- H. Qu, R. Yan, X. Shu, H. Gao, P. Huang, and G.-S. Xie, "MVP-shot: Multi-velocity progressive-alignment framework for few-shot action recognition," 2024, *arXiv:2405.02077*.
- J. Xing et al., "MA-FSAR: Multimodal adaptation of CLIP for few-shot action recognition," 2023, *arXiv:2308.01532*.
- R. Goyal et al., "The 'something something' video database for learning and evaluating visual common sense," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Sep. 2017, pp. 5842–5850.
- Q. Wang, J. Du, K. Yan, and S. Ding, "Seeing in flowing: Adapting CLIP for action recognition with motion prompts learning," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 5339–5347.
- X. Wang, H. Zhang, S. Zhang, C. Gao, Y. Shao, and N. Sang, "Context-aware proposal network for temporal action detection," 2022, *arXiv:2206.09082*.
- M. Jain, H. Jégou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2555–2562.
- J. Wu, T. Zhang, Z. Zhang, F. Wu, and Y. Zhang, "Motion-modulated temporal fragment alignment network for few-shot action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9151–9160.
- L. Wang, Z. Tong, B. Ji, and G. Wu, "TDN: Temporal difference networks for efficient action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1895–1904.
- K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16816–16825.
- S. Nag et al., "Multi-modal few-shot temporal action detection," 2022, *arXiv:2211.14905*.
- S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang, "Zero-shot temporal action detection via vision-language prompting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jan. 2022, pp. 681–697.
- W. Zhang, C. Wan, T. Liu, X. Tian, X. Shen, and J. Ye, "Enhanced motion-text alignment for image-to-video transfer learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 18504–18515.
- C. Careaga, B. Hutchinson, N. Hodas, and L. Phillips, "Metric-based few-shot learning for video action recognition," 2019, *arXiv:1909.09602*.
- S. K. Dwivedi, V. Gupta, R. Mitra, S. Ahmed, and A. Jain, "ProtoGAN: Towards few shot learning for action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1308–1316.
- X. Zhu, A. Toisoul, J.-M. Perez-Rua, L. Zhang, B. Martinez, and T. Xiang, "Few-shot action recognition with prototype-centered attentive learning," 2021, *arXiv:2101.08085*.

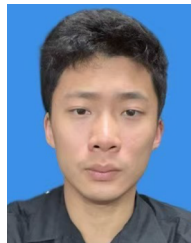
- [35] A. Thatipelli, S. Narayan, S. Khan, R. M. Anwer, F. S. Khan, and B. Ghanem, "Spatio-temporal relation modeling for few-shot action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19958–19967.
- [36] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, "TEA: Temporal excitation and aggregation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2020, pp. 909–918.
- [37] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Aug. 2017, pp. 5533–5541.
- [38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Apr. 2015, pp. 4489–4497.
- [39] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.
- [40] S. Jiang, Y. Qi, H. Zhang, Z. Bai, X. Lu, and P. Wang, "D3D: Dual 3-D convolutional network for real-time action recognition," *IEEE Trans. Ind. Informat.*, vol. 17, no. 7, pp. 4584–4593, Jul. 2021.
- [41] Y. Yi et al., "High performance gesture recognition via effective and efficient temporal modeling," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1003–1009.
- [42] B. Jiang, M. Wang, W. Gan, W. Wu, and J. Yan, "STM: Spatiotemporal and motion encoding for action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2000–2009.
- [43] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.
- [44] B. Mohammadi, Y. Hong, Y. Qi, Q. Wu, S. Pan, and Q. Shi, "Augmented commonsense knowledge for remote object grounding," in *Proc. Assoc. Advancement Artif. Intell. (AAAI)*, vol. 38, Mar. 2024, pp. 4269–4277.
- [45] D. An et al., "BEVBert: Multimodal map pre-training for language-guided navigation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 2737–2748.
- [46] V. M. H. Phan et al., "Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 11492–11501.
- [47] W. Chen, C. Si, Z. Zhang, L. Wang, Z. Wang, and T. Tan, "Semantic prompt for few-shot image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jan. 2023, pp. 23581–23591.
- [48] B. Ni et al., "Expanding language-image pretrained models for general video recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jan. 2022, pp. 1–18.
- [49] J. Xing, M. Wang, B. Mu, and Y. Liu, "Revisiting the spatial and temporal modeling for few-shot action recognition," in *Proc. Assoc. Advancement Artif. Intell. (AAAI)*, Jan. 2023, pp. 3001–3009.
- [50] L. Wang et al., "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.
- [51] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, Sep. 2022.
- [52] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 1126–1135.
- [53] L. Zhu and Y. Yang, "Compound memory networks for few-shot video classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 751–766.
- [54] S. Zhang, J. Zhou, and X. He, "Learning implicit temporal alignment for few-shot video classification," in *Proc. Thirtieth Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1309–1315.
- [55] S. Li et al., "TA2N: Two-stage action alignment network for few-shot action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 1404–1411.
- [56] Y. Huang, L. Yang, and Y. Sato, "Compound prototype matching for few-shot action recognition," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel, Cham, Switzerland: Springer, Oct. 2022, pp. 351–368.
- [57] K. D. Nguyen, Q.-H. Tran, K. Nguyen, B.-S. Hua, and R. Nguyen, "Inductive and transductive few-shot video classification via appearance and temporal alignments," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Cham, Switzerland: Springer, Oct. 2022, pp. 471–487.
- [58] J. Xing et al., "Boosting few-shot action recognition with graph-guided hybrid matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1740–1750.
- [59] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.
- [60] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.
- [61] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.
- [62] H. Zhang, L. Zhang, X. Qi, H. Li, P. H. Torr, and P. Koniusz, "Few-shot action recognition with permutation-invariant attention," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 525–542.
- [63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [64] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2009, pp. 248–255.
- [65] H. Xu et al., "VideoCLIP: Contrastive pre-training for zero-shot video-text understanding," 2021, *arXiv:2109.14084*.
- [66] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.



Pengfei Fang received the M.E. degree from The Australian National University in 2017 and the Ph.D. degree from The Australian National University and DATA61-CSIRO in 2022. He is currently a Professor at the School of Computer Science and Engineering, Southeast University (SEU), China. He is also a member of the PALM Laboratory. Before joining SEU, he was a Post-Doctoral Fellow at Monash University in 2022. His research interests include computer vision and machine learning.



Qiang Xu received the B.Sc. degree from Nanjing Tech University, Nanjing, China. He is currently pursuing the master's degree with the School of Computer Science and Engineering, Southeast University, Nanjing. His research interests include computer vision and machine learning.



Zixuan Lin is currently pursuing the B.Sc. degree in mechanical engineering with Xi'an Jiaotong University, Xi'an, China. His research interests include machine learning.



Hui Xue (Member, IEEE) received the B.Sc. degree in mathematics from Nanjing Normal University, Nanjing, China, in 2002, and the M.Sc. degree in mathematics and the Ph.D. degree in computer application technology from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, in 2005 and 2008, respectively. She is currently a Professor with the School of Computer Science and Engineering, Southeast University, Nanjing. Her research interests include pattern recognition and machine learning.