Task-Specific Distance Correlation Matching for Few-Shot Action Recognition

Fei Long[1*], **Yao Zhang**[1*], **Jiaming Lv**[1], **Jiangtao Xie**[1], **Peihua Li**[1†]

[1]School of Information and Communication Engineering, Dalian University of Technology
longfei121@mail.dlut.edu.cn, zyemail@mail.dlut.edu.cn, Peihua Li@dlut.edu.cn

Abstract

Few-shot action recognition (FSAR) has recently made notable progress through set matching and efficient adaptation of large-scale pre-trained models. However, two key limitations persist. First, existing set matching metrics typically rely on cosine similarity to measure inter-frame linear dependencies and then perform matching with only instance-level information, thus failing to capture more complex patterns such as nonlinear relationships and overlooking task-specific cues. Second, for efficient adaptation of CLIP to FSAR, recent work performing fine-tuning via skip-fusion layers (which we refer to as side layers) has significantly reduced memory cost. However, the newly introduced side layers are often difficult to optimize under limited data conditions. To address these limitations, we propose TS-FSAR, a framework comprising three components: (1) a visual Ladder Side Network (LSN) for efficient CLIP fine-tuning; (2) a metric called Task-Specific Distance Correlation Matching (TS-DCM), which uses α-distance correlation to model both linear and nonlinear inter-frame dependencies and leverages a task prototype to enable task-specific matching; and (3) a Guiding LSN with Adapted CLIP (GLAC) module, which regularizes LSN using the adapted frozen CLIP to improve training for better α-distance correlation estimation under limited supervision. Extensive experiments on five widely-used benchmarks demonstrate that our TS-FSAR yields superior performance compared to prior state-of-the-arts.

1 Introduction

Action recognition, as a fundamental task in visual learning, has attracted widespread attention and achieved remarkable success in recent years (Wang et al. 2016; Carreira and Zisserman 2017a; Wang et al. 2018; Zhou et al. 2018; Lin, Gan, and Han 2019; Arnab et al. 2021; Park, Lee, and Sohn 2023; Qian, Ding, and Lin 2024). However, training for this task typically requires large-scale labeled video datasets, which are costly to collect and often impractical in real-world scenarios. To overcome this limitation, few-shot action recognition (FSAR) seeks to recognize previously unseen classes from only limited labeled examples, and has thus emerged as an active and growing research area within the community (Zhu and Yang 2018; Cao et al. 2020; Thatipelli et al. 2022; Kumar et al. 2024; Lee et al. 2025).

In few-shot action recognition, most existing methods focus on designing more effective metrics and exploring efficient adaptation from large-scale pre-trained models—such as ImageNet-1K pretrained models (He et al. 2016; Dosovitskiy et al. 2021) or CLIP (Radford et al. 2021). To obtain a more discriminative metric, it is particularly important to consider how the distance between query and support prototypes is computed. In contrast to early methods (Cao et al. 2020) that formulate the distance computation as a temporal alignment problem, several recent methods reformulate this as a matching problem between two sets of features, employing techniques such as Hausdorff distance (Wang et al. 2022) or optimal transport (Wu et al. 2022; Li et al. 2025) to compute the distance. Despite effectiveness, they still suffer from two primary limitations. First, they typically rely on cosine similarity to construct a distance matrix that captures inter-frame relationships between query and support. However, cosine similarity is approximately equivalent to Pearson Correlation Coefficient (Zhelezniak et al. 2019), which can only capture linear relationships and thus fails to model more complex correlations such as nonlinear dependencies. Second, they perform matching based on instance-level information without considering task-specific cues, whose effectiveness has been demonstrated in prior works (Wu et al. 2022; Wang et al. 2022; Cao et al. 2024). This limits their ability to achieve more accurate matching. To enable efficient adaptation of CLIP in FSAR, recent works (Pei et al. 2025; Xing et al. 2025; Li et al. 2025) have introduced parameter-efficient adapters instead of full fine-tuning (Wang et al. 2024). However, these approaches still require backpropagation through the backbone, leading to considerable GPU memory consumption. In contrast, EMP-Net (Wu et al. 2024) proposes to fine-tune the newly introduced skip-fusion layers that take as input the activation features from the frozen CLIP, thereby avoiding backbone backpropagation and reducing memory usage. This design shares a similar principle with Ladder Side-Tuning (Zhang et al. 2020). Despite being memory-efficient, optimizing skip-fusion layers under limited data remains challenging, especially on static datasets that depend heavily on pre-trained knowledge.

[*]These authors contributed equally.
[†]Corresponding Author.

Inspired by the observations above, we propose TS-FSAR, which introduces a novel metric that matches two set of features in a task-specific manner, and fine-tunes the CLIP vision encoder via a ladder side network (LSN) guided by the frozen CLIP equipped with an adapter. In particular, the metric we propose can be decomposed into two components. First, to capture comprehensive inter-frame dependencies between query and support, we adopt α-distance correlation (Székely and Rizzo 2009), an extension of distance correlation (Székely and Rizzo 2009) known for capturing both linear and nonlinear relationships, to compute a inter-frame α-correlation matrix for each query-support pair. Second, to achieve task-specific matching, we employs a learnable generator which takes a query-specific task prototype as input to produce an matching matrix, which encodes the relative importance of relationships between different frames of the query and support videos. Then, the final similarity score is obtained by taking the inner product of the inter-frame α-correlation matrix and the matching matrix. To improve the training of the LSN under limited data for reliable α-distance correlation estimation, we obtain a distribution by applying softmax over the α-distance correlations between LSN features and different class prototypes, and align it with the adapted frozen CLIP's output distribution. To validate the effectiveness of our proposed TS-FSAR, we evaluated it on five standard datasets. The experimental results demonstrate that our method achieves superior performance compared to prior methods. Our contributions can be summarized as follows:

- We propose a novel metric, termed Task-specific Distance Correlation Matching (TS-DCM), for few-shot action recognition. It leverages α-distance correlations to measure inter-frame relationships and employs a query-specific task prototype to perform task-specific matching.
- We introduce a Guiding LSN with Adapted CLIP (GLAC) module to guide the LSN using the output distribution of the adapted frozen CLIP, which helps improve the training of the LSN under limited data conditions, thereby contributing to more reliable α-distance correlation estimation based on its features.
- Our proposed TS-FSAR achieves leading performance across several few-shot action recognition benchmarks, with particularly significant improvements on temporally challenging datasets such as SSv2-Full.

2 Related Work

Few-shot Action Recognition In few-shot action recognition, a series of methods focus on enhancing video representations to improve generalization. AMeFu-Net (Fu et al. 2020) leverages depth information through an adaptive normalization module with temporal asynchronization to enrich cross-modal representations. TRX (Perrett et al. 2021) builds query-specific class prototypes using cross-attention across ordered video sub-sequences. TA^{2}N (Li et al. 2022) introduces a two-stage alignment framework to mitigate temporal and spatial misalignment via temporal transformation and action coordination. MoLo (Wang et al. 2023) proposes motion-augmented long-short contrastive learning to jointly capture long-range temporal dependencies and motion cues. Beyond these representation-oriented works, another branch of research aims to design more discriminative metrics. OTAM (Cao et al. 2020) employs a variant of Dynamic Time Warping to align query–support sequences. HyRSM (Wang et al. 2022) formulates sequence distance computation as a set-matching problem and introduces bidirectional Mean Hausdorff matching. Following this paradigm, MTFAN (Wu et al. 2022) and TSAM (Li et al. 2025) compute the distance between query and support features using Optimal Transport. Our proposed metric differs from existing methods in two key aspects. First, instead of using cosine similarity to measure inter-frame relations, we adopt α-distance correlation to achieve more comprehensive modeling of inter-frame dependencies. Second, unlike these metrics that follow a task-agnostic paradigm, our proposed TS-DCM perform matching between query and support videos in a task-specific manner, resulting in more accurate matching.

Side-Tuning for Video Understanding Recently, efficiently adapting CLIP to domain-specific tasks has gained growing attention. Among existing methods, side-tuning (Zhang et al. 2020; Sung, Cho, and Bansal 2022) provides a memory-efficient solution by attaching a lightweight side network to the frozen backbone, and has been widely adopted in recent studies, including those on video understanding. EVL (Lin et al. 2022) introduces an efficient video recognition framework using frozen CLIP features, where a lightweight Transformer decoder and local temporal modules are used to capture spatial and temporal cues; STAN (Liu et al. 2023) introduces a spatial-temporal auxiliary network with a branched architecture, which incorporates decomposed spatial-temporal modules to effectively contextualize multilevel CLIP features for video tasks; EMP-Net (Wu et al. 2024) leverages skip-fusion layers to integrate multi-stage CLIP features and performs multi-level post-reasoning in few-shot action recognition. Unlike these methods that focus on designing different side network architectures, we directly adopt a simple Ladder Side Network (Sung, Cho, and Bansal 2022) (LSN) for finetuning. During training, the output distribution of adapted frozen CLIP is used to guide the LSN, enabling more effective optimization under limited data conditions.

3 Methodology

3.1 Problem Formulation

Few-shot action recognition (FSAR) aims to learn a model that can generalize to unseen action categories using only a few labeled examples. A FSAR dataset is typically divided into three disjoint subsets: $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{test}}$, each containing non-overlapping classes. To align with the evaluation scenario, FSAR models are typically trained in an episodic manner. Specifically, each episode comprises a support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N_\mathcal{S}}$ and a query set $\mathcal{Q} = \{(x_i, y_i)\}_{i=1}^{N_\mathcal{Q}}$. The support set consists of K labeled videos per class from N classes, forming an N-way K-shot task with $N_\mathcal{S} = N \times K$ samples in total. The query set, denoted by $\mathcal{Q}$, includes $N_\mathcal{Q}$ samples to be classified.

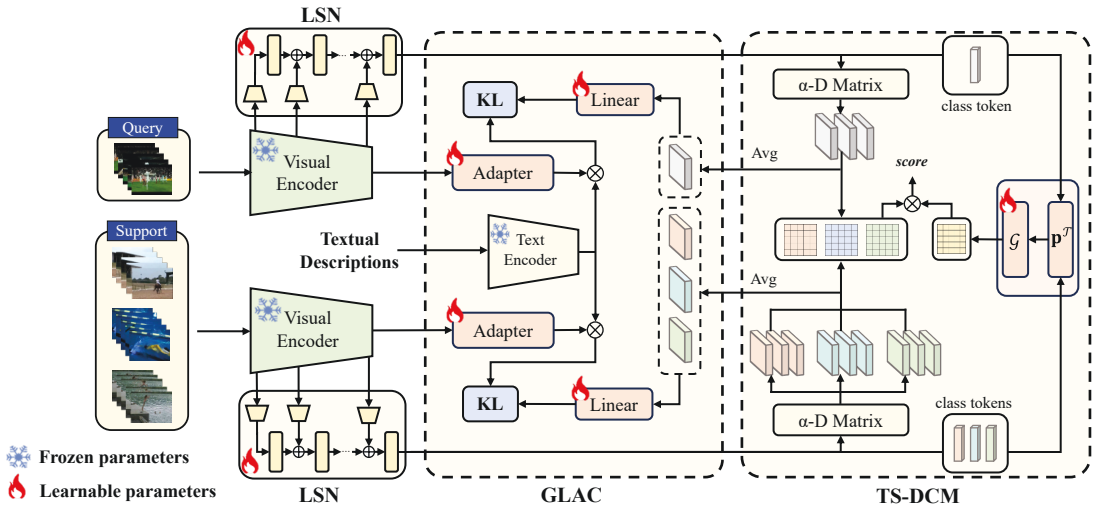

Figure 1: Illustration of the proposed TS-FSAR framework, which comprises three components: (1) a Ladder Side Network (LSN) for memory-efficient fine-tuning of the CLIP visual encoder, (2) a metric named Task-Specific Distance Correlation Matching (TS-DCM) that leverages α-distance correlation and a task prototype for more accurate query-support matching, and (3) a Guiding LSN with Adapted CLIP (GLAC) module that enhances LSN training under limited data to enable more reliable distance correlation estimation.

3.2 Overview of TS-FSAR

As illustrated in Figure 1, the proposed TS-FSAR framework comprises three main components: a Ladder Side Network (LSN), a metric called Task-Specific Distance Correlation Matching (TS-DCM), and a Guiding LSN with Adapted CLIP (GLAC) module. In our framework, both query and support videos are processed in parallel, sharing all learnable parameters across the LSN, adapter, and linear layer.

Take 1-shot for an example. Given a task, both query and support videos are first fed into the LSN to extract output features. Then, these features are processed by the TS-DCM to perform task-specific matching. It begins by computing the frame-level α-distance (α-D) matrices for the query and support videos independently, followed by deriving an inter-frame α-distance correlation matrix for each query-support pair, capturing both linear and nonlinear dependencies between their frames. Then, a query-specific prototype is constructed by the class tokens of the support and query videos. This prototype is then passed into a learnable generator to produce a matching matrix that encodes the relative importance of inter-frame relationships between the query and support. Finally, the similarity scores for the query is obtained by computing the inner product between the matching matrix and the inter-frame α-distance correlation matrices.

To improve the training of LSN under limited data for more reliable α-distance correlation estimation, we introduce the GLAC module. First, both query and support videos are processed by frozen CLIP visual encoder, followed by an adapter that helps adapt CLIP to the video domain. The adapted visual features are then aligned with text embeddings to obtain a guidance distribution. Meanwhile, we average the frame-level α-D matrices for each video to obtain video-level representation, and then obtain a distribution by applying softmax to the α-distance correlations between the representation and learnable class prototypes implemented via a linear layer. The training of LSN is then guided by minimizing the KL divergence between these two distributions.

3.3 Ladder Side Network

Ladder Side Tuning (LST) (Sung, Cho, and Bansal 2022) is a memory-efficient fine-tuning strategy that introduces a lightweight and separate Ladder Side Network (LSN) alongside the backbone. The LSN receives dimension-reduced hidden features from the backbone via shortcut connections as input. To enable efficient adaptation, we employ the LSN to fine-tune the vision encoder of CLIP.

During training, for each video, we align the visual outputs of the LSN with the corresponding text embeddings. Given a video with T frames, the visual token embeddings of the t-th frame produced by the LSN can be denoted as $\mathbf{V}^t = [\mathbf{v}_0^t, \cdots, \mathbf{v}_P^t] \in \mathbb{R}^{(P+1)\times d}$, where P is the number of patch tokens, and d denotes the feature dimension. Next, we project the class token $\mathbf{v}_0^t$ through a linear layer to match the dimension of the text embeddings, yielding $\hat{\mathbf{v}}_0^t \in \mathbb{R}^{d'}$. We then average the class tokens $\hat{\mathbf{v}}_0^t$ from all frames to obtain the video-level representation $\widetilde{\mathbf{v}}$. Let $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_C]$ denote the text embeddings obtained from text encoder, where $\mathbf{w}_i \in \mathbb{R}^{d'}$ represents the embedding of the i-th class description, C denotes the number of classes in $\mathcal{D}_{\text{train}}$. We then compute the cosine similarity between $\widetilde{\mathbf{v}}$ and $\mathbf{w}_i$, followed by a softmax operation to obtain the prediction. The resulting prediction is then used to define the training loss via cross-entropy, denoted as $\mathcal{L}_{\text{LSN}}$. *More implementation details of the LSN can be found in Appendix Sec.A.*

3.4 Task-Specific Distance Correlation Matching

Our proposed TS-DCM can be decomposed into two components: Inter-Frame α-Distance Correlation (IF-D$^\alpha$C) and Task-Specific Matching (TSM). IF-D$^\alpha$C computes inter-frame correlation matrices between support and query videos using α-distance correlation, capturing more comprehensive inter-frame dependencies. Then, TSM takes the task prototype as input to a learnable generator, which produces a matching matrix for performing task-specific matching between the query and support.

α-Distance Correlation α-distance correlation is an extension of standard distance correlation (Székely and Rizzo 2009), which is known for effectively modeling both linear and nonlinear relationships between random variables. The α-distance covariance between two random variables $\mathbf{X}$ and $\mathbf{Y}$ is defined as α-weighted L^2 distance between their joint characteristic function $\varphi_{\mathbf{X},\mathbf{Y}}(t,s)$ and the product of their marginal characteristic functions $\varphi_{\mathbf{X}}(t)$, $\varphi_{\mathbf{Y}}(s)$:

$$\mathrm{DCov}^{2(\alpha)}(\mathbf{X},\mathbf{Y}) = \|\varphi_{\mathbf{X},\mathbf{Y}}(t,s) - \varphi_{\mathbf{X}}(t)\varphi_{\mathbf{Y}}(s)\|_\alpha^2 \quad (1)$$

The exponent parameter $\alpha \in (0,2)$ serves to modulate the sensitivity to dependencies of varying scales. The α-distance correlation is defined as the normalized version of α-distance covariance, and serves to quantify the degree of dependence between two random variables.

In the discrete case, one can follow the procedure in (Székely and Rizzo 2009) to compute the empirical α-distance correlation. Given two random variables $\mathbf{X}$ and $\mathbf{Y}$ with m independent and identically distributed observations $\{(\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_m,\mathbf{y}_m)\}$, we first compute the α-th power of pairwise Euclidean distances to obtain the matrices $\widehat{\mathbf{A}} = (\widehat{a}_{kl})$ and $\widehat{\mathbf{B}} = (\widehat{b}_{kl})$, as follows:

$$\widehat{a}_{kl} = \|\mathbf{x}_k - \mathbf{x}_l\|^\alpha, \widehat{b}_{kl} = \|\mathbf{y}_k - \mathbf{y}_l\|^\alpha \quad (2)$$

By applying standard double centering to matrix $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$, we obtain the centered α-distance matrices (referred to as α-D matrix) for $\mathbf{X}$ and $\mathbf{Y}$, denoted as $\mathbf{A}$ and $\mathbf{B}$, respectively. Next, the α-distance covariance between $\mathbf{X}$ and $\mathbf{Y}$ can be obtained as:

$$\mathrm{DCov}^{2(\alpha)}(\mathbf{X},\mathbf{Y}) = \frac{1}{m^2}\mathtt{tr}(\mathbf{AB}) \quad (3)$$

Then, the α-distance correlation is defined as:

$$\mathrm{DCorr}^{2(\alpha)}(\mathbf{X},\mathbf{Y}) = \frac{\mathtt{tr}(\mathbf{AB})}{\sqrt{\mathtt{tr}(\mathbf{AA})}\sqrt{\mathtt{tr}(\mathbf{BB})}} \quad (4)$$

Inter-Frame α-Distance Correlation Let $\mathbf{V}_{\mathcal{S}}^i \in \mathbb{R}^{(P+1)\times d}$ and $\mathbf{V}_{\mathcal{Q}}^j \in \mathbb{R}^{(P+1)\times d}$ denote the features of the i-th frame from a support video and the j-th frame from a query video, respectively. By treating each column of $\mathbf{V}_{\mathcal{S}}^i$ and $\mathbf{V}_{\mathcal{Q}}^j$ as an observation of random vectors $\mathbf{X}$ and $\mathbf{Y}$, we compute the corresponding α-D matrices, denoted as $\mathbf{A}^i$ and $\mathbf{B}^j$. Then, the α-Distance Correlation m_{ij} between the i-th frame of the support video and the j-th frame of the query video can be computed using Eq. (4). Based on this, the inter-frame α-Distance Correlation matrix is formed as $\mathbf{M}^{\text{IF-D}^\alpha\text{C}} = (m_{ij}) \in \mathbb{R}^{T\times T}$.

Task-Specific Matching The matrix $\mathbf{M}^{\text{IF-D}^\alpha\text{C}}$ encodes inter-frame correlations between the support and query videos. The key to leveraging these correlations lies in designing an appropriate matching matrix that captures the relative importance between query and support frames. To achieve this, a task prototype is fed into a learnable generator to produce the matching matrix $\mathbf{M}^{\text{task}}$, enabling flexible and task-specific matching.

Let $\widetilde{\mathbf{v}}_i^{\mathcal{S}}$ and $\widetilde{\mathbf{v}}^{\mathcal{Q}}$ denote the frame-wise averaged class token produced by LSN for the support video $x_i \in \mathcal{S}$ and a query video $x \in \mathcal{Q}$, respectively. Then we design a query-specific task prototype , which can be computed as :

$$\mathbf{p}^{\mathcal{T}} = \widetilde{\mathbf{v}}^{\mathcal{Q}} + \frac{1}{N_{\mathcal{S}}}\sum_{x_i\in\mathcal{S}} \widetilde{\mathbf{v}}_i^{\mathcal{S}} \quad (5)$$

where $N_{\mathcal{S}}$ is the number of videos in the entire support set. Besides average-based fusion, we also explore alternative strategies such as concatenation and cross-attention, which will be discussed in Sec. 4.4.

After getting the task prototype, we employ a learnable linear layer as the generator $\mathcal{G}(\cdot)$ to produce the task-specific matching matrix $\mathbf{M}^{\text{task}} \in \mathbb{R}^{T\times T}$:

$$\mathbf{M}^{\text{task}} = \mathcal{G}(\mathbf{p}^{\mathcal{T}}) \quad (6)$$

And then, the similarity score between the query video and the support video can be obtained as:

$$score = \langle \mathbf{M}^{\text{task}}, \mathbf{M}^{\text{IF-D}^\alpha\text{C}} \rangle \quad (7)$$

where $\langle\cdot\rangle$ donote the inner product.

For a given query video, we compute the scores by matching it with the support video of each class. Applying softmax to these scores yields a prediction probability vector $\mathbf{s} = [s_1, s_2, \ldots, s_N]$. We then optimize the model by computing the cross-entropy loss between the ground-truth label and the prediction vector $\mathbf{s}$, denoted as $\mathcal{L}_{\text{TS-DCM}}$.

3.5 Guiding LSN with Adapted CLIP

Although finetuning with LSN is memory-efficient, optimizing a number of newly introduced layers with limited samples remains highly challenging. This directly influences the estimation of inter-frame α-distance correlation derived from the output features of LSN. Motivated by this, we propose to guide the training of the LSN through alignment with the output distribution of the adapted frozen CLIP.

For a given video, we average the frame-wise α-D matrices to obtain the video-level representation $\widetilde{\mathbf{A}}_{\alpha-\mathrm{D}} \in \mathbb{R}^{d\times d}$. Then, to better adapt the estimation of the α-distance correlation in TS-DCM, we initialize learnable weight matrix $\widetilde{\mathbf{W}}_i$ for each class i as its α-D matrix prototype. These weight matrices can be implemented using a linear layer. Then, the prediction is computed by taking the inner product between $\widetilde{\mathbf{A}}_{\alpha\text{-D}}$ and $\widetilde{\mathbf{W}}_i$, followed by a softmax operation to produce the predicted probability vector $\mathbf{p} = [p_1, p_2, ..., p_C]$.

To improve the guidance, we employ an adapter composed of a standard multi-head self-attention (MHSA) to help CLIP adapt to the video domain, where frame-level CLS tokens are fed into the adapter to model inter-frame

dependencies. Subsequently, the class tokens from different frames produced by the adapter are averaged to form a video-level representation, denoted by $\widetilde{e}$. Then, we calculate the cosine similarity between $\widetilde{e}$ and the textual features output by text encoder, and apply a softmax function to obtain the guidance vector $\mathbf{q} = [q_1, q_2, ..., q_C]$. Then, we employ the Kullback–Leibler (KL) divergence to guide the learning process, and the loss is defined as follows:

$$\mathcal{L}_{\text{GLAC-KL}} = \text{KL}(\mathbf{p} \parallel \mathbf{q}) = \sum_{i=1}^{C} p_i \log \frac{p_i}{q_i} \tag{8}$$

Furthermore, ground-truth labels are employed to supervise both the CLIP and LSN branches through cross-entropy loss, thereby improving their individual training processes.

$$\mathcal{L}_{\text{GLAC-CE}} = -\sum_{i=1}^{C} y_i \log(p_i) + (-\sum_{i=1}^{C} y_i \log(q_i)) \tag{9}$$

Finally, the total loss for GLAC module can be defined as:

$$\mathcal{L}_{\text{GLAC}} = \mathcal{L}_{\text{GLAC-KL}} + \mathcal{L}_{\text{GLAC-CE}} \tag{10}$$

3.6 Training Loss

The training loss of our TS-FSAR is composed of the three components described above: the vision-language alignment loss for the LSN, the TS-DCM loss, and the GLAC loss. Accordingly, the total loss can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{LSN}} + \lambda_1 \mathcal{L}_{\text{TS-DCM}} + \lambda_2 \mathcal{L}_{\text{GLAC}} \tag{11}$$

where λ_1 and λ_2 are weights tuned on the validation set.

4 Experiments

4.1 Datasets

We evaluate our method on five commonly used benchmarks: SSv2-Full (Goyal et al. 2017), SSv2-Small (Goyal et al. 2017), Kinetics-100 (Carreira and Zisserman 2017b), UCF101 (Soomro, Zamir, and Shah 2012), and HMDB51 (Kuehne et al. 2011). *Please refer to Appendix Sec.B.1 for details about the dataset.*

4.2 Implementation

Following previous works (Wang et al. 2024; Wu et al. 2024; Li et al. 2025), we implement our framework using CLIP ViT-B/16 as the visual backbone and uniformly sample 8 frames to construct the input sequence for each video. The dimension of the LSN is set to 256, and the number of layers is aligned with the visual encoder, i.e., 12. Following TSAM (Li et al. 2025), we employ large language model to generate class-specific descriptions. During training, we use the AdamW optimizer with a weight decay of 0.1 and adopt a cosine learning rate schedule. The learning rate is initialized to 2e-4 for SSv2-Full, SSv2-Small, and HMDB51, and to 1e-4 in the 5-shot and 2e-4 in the 1-shot for Kinetics-100 and UCF101. We sample 50,000 training episodes for SSv2-Full, and 10,000 episodes for all other datasets. We report the average accuracy over 10,000 episodes. *More implementation details are provided in Appendix Sec.B.2.*

4.3 Comparison with State-of-the-Art Methods

We compare TS-FSAR with recent state-of-the-art approaches across both temporally-dependent and spatially-dependent FSAR benchmarks, as summarized in Table 1.

Temporally-dependent datasets On SSv2-Small, TS-FSAR performs comparably to the state-of-the-art method TSAM in 1-shot and outperforms existing methods by 1% in 5-shot. On SSv2-Full, TS-FSAR significantly outperforms prior arts by 8.4% in 1-shot and 1.6% in 5-shot, demonstrating its superior ability to model complex temporal dependencies in few-shot scenarios.

Spatially-dependent datasets On HMDB51, UCF101, and Kinetics-100, our TS-FSAR performs slightly better or comparably to the previous leading method TSAM, except for a minor drop in 5-shot on Kinetics-100. We suspect that this is due to the relatively weak temporal dynamics of these datasets. Compared to SSv2-Full and SSv2-Small, these datasets are more aligned with CLIP's pretraining distribution, which consists of static images lacking temporal structure, and thus rely more heavily on CLIP's pretrained weights. As a result, fine-tuning the LSN on their limited training set becomes more challenging. Compared to EMP-Net, which also employs a side network, our method benefits from the GLAC module, achieving 1.1%$\sim$8.2% performance gains. However, this design still cannot fully resolve the optimization challenge under such conditions.

Why the large gain on SSv2-Full? Our method shows a significantly stronger performance on SSv2-Full, which we attribute to two main factors. First, SSv2-Full exhibits fine-grained temporal variations absent in K100, UCF101, and HMDB51, where our task-specific matching effectively captures such detailed temporal dynamics. Second, its considerably larger base set—about 10 times that of SSv2-Small and other datasets—provides better supervision for LSN training and leads to more reliable α-DC estimation. Together, these factors explain the larger gain on SSv2-Full.

4.4 Ablation Study

To better understand the design choices and individual contributions of our proposed TS-FSAR framework, we perform ablation studies on SSv2-Full and HMDB51.

Ablation on Main Components We present a comprehensive ablation study to quantify the contribution of each main component within the TS-FSAR framework, as detailed in Table 2. The analysis begins with the zero-shot CLIP baseline, which exhibits limited performance, achieving 37.0% on SSv2-Full and 75.9% on HMDB51. By introducing the LSN, we observe a marked increase in accuracy. Next, to better understand our proposed TS-DCM metric, we perform ablation by decoupling it into two components: IF-D$^{\alpha}$C and TSM. Based on the LSN, introducing IF-D$^{\alpha}$C further improves performance by 4.3%$\sim$4.6% across all benchmarks, demonstrating the benefit of fully capturing inter-frame dependencies. Further incorporating the TSM module yields an additional gain of 1.1%$\sim$2.4%, highlighting the advantage of introducing task-specific information when performing matching. Finally, the incorporation of GLAC consistently enhances performance, with especially notable

Method	Backbone	SSv2-Full		SSv2-Small		HMDB51		UCF101		Kinetics	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
OTAM (Cao et al. 2020)	IN-RN50	42.8	52.3	36.4	48.0	54.5	68.0	79.9	88.9	73.0	85.8
TRX (Perrett et al. 2021)	IN-RN50	42.0	64.6	36.0	56.7	54.9	75.6	81.0	96.1	65.1	85.9
STRM (Thatipelli et al. 2022)	IN-RN50	43.1	68.1	37.1	55.3	57.6	77.3	82.7	96.9	65.1	86.7
HyRSM (Wang et al. 2022)	IN-RN50	54.3	69.0	40.6	56.1	60.3	76.0	83.9	94.7	73.7	86.1
HCL (Zheng, Chen, and Jin 2022)	IN-RN50	47.3	64.9	38.7	55.4	59.1	76.3	82.6	94.5	73.7	85.8
Nguyen (Nguyen et al. 2022)	IN-RN50	43.8	61.1	–	–	59.6	76.9	84.9	95.9	74.3	87.4
SloshNet (Xing et al. 2023a)	IN-RN50	46.5	68.3	–	–	59.4	77.5	86.0	97.1	70.4	87.0
GgHM (Xing et al. 2023b)	IN-RN50	54.5	69.2	–	–	61.2	76.9	85.2	96.3	74.9	87.4
TEAM (Lee et al. 2025)	IN-RN50	–	–	–	–	62.8	78.4	87.2	96.2	75.1	88.2
CLIP-FSAR (Wang et al. 2024)	CLIP-ViT-B/16	62.1	72.1	54.6	61.8	77.1	87.7	97.0	99.1	94.8	95.4
EMP-Net (Wu et al. 2024)	CLIP-ViT-B/16	63.1	73.0	57.1	65.7	76.8	85.8	94.3	98.2	89.1	93.5
MVP-shot (Qu et al. 2025)	CLIP-ViT-B/16	-	-	55.4	62.0	77.0	88.1	96.8	99.0	91.0	95.1
MA-FSAR (Xing et al. 2025)	CLIP-ViT-B/16	63.3	72.3	59.1	64.5	83.4	87.9	97.2	99.2	95.7	96.0
D^2ST-Adapter (Pei et al. 2025)	CLIP-ViT-B/16	66.7	81.9	55.0	69.3	77.1	88.2	96.4	99.1	89.3	95.5
TSAM (Li et al. 2025)	CLIP-ViT-B/16	65.8	74.6	**60.5**	66.7	84.5	**88.9**	98.3	99.3	96.2	97.1
TS-FSAR (Ours)	CLIP-ViT-B/16	**75.1**	**83.5**	**60.5**	**70.3**	**85.0**	**88.9**	**98.7**	**99.3**	**96.3**	96.6

Table 1: Performance comparison with state-of-the-art methods on standard benchmarks. IN denotes ImageNet.

LSN	IF-D$^\alpha$C	TSM	GLAC	SSv2-Full 1-shot	SSv2-Full 5-shot	HMDB51 1-shot	HMDB51 5-shot
				37.0	37.0	75.9	75.9
✓				67.1	77.2	77.7	79.6
✓	✓			71.4	81.7	82.1	84.2
✓	✓	✓		73.8	82.8	83.4	85.6
✓	✓	✓	✓	75.1	83.5	85.0	88.9

Table 2: Ablation on key components of TS-FSAR

improvements on HMDB51, validating our hypothesis that insufficient training of the LSN has a greater impact on static datasets. Furthermore, on SSv2-Full, removing the adapter in GLAC leads to a 3.6% performance drop.

About the IF-D$^\alpha$C We employ α-DC to model inter-frame correlations, which can capture complex (nonlinear) dependencies. Similar alternatives include DC used in DeepBDC (Xie et al. 2022) and the kernel-based HSIC (Gretton et al. 2005). To compare them fairly, we replaced only the inter-frame similarity metric under identical settings. As shown in the Figure 3, nonlinear measures (α-DC, DC, HSIC) consistently outperform Cosine Similarity (CS), while α-DC achieves the best performance owing to its more robust nonlinear modeling. As noted by (Leyder, Raymaekers, and Rousseeuw 2024), the coefficient α acts as an empirical hyperparameter that governs the trade-off between robustness and sensitivity. Hence, we further investigate the impact of α (Eq. (2)) to analyze its influence on performance. As shown in Figure 2, $\alpha = 1.2$ achieved the highest 1-shot accuracy on SSv2-Full, but $\alpha = 0.8$ yielded the best overall results. Consequently, we selected $\alpha = 0.8$ as the default throughout the paper.

Effect of Different Task Prototypes To evaluate the ef-

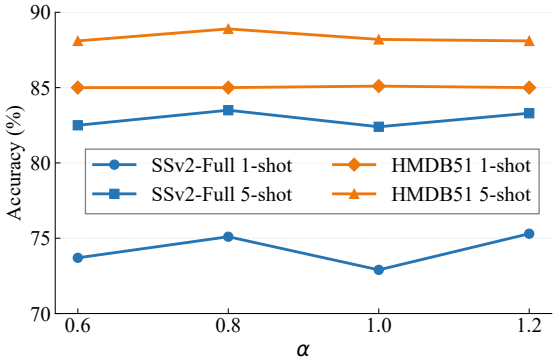

Figure 2: The impact of α in IF-D$^\alpha$C

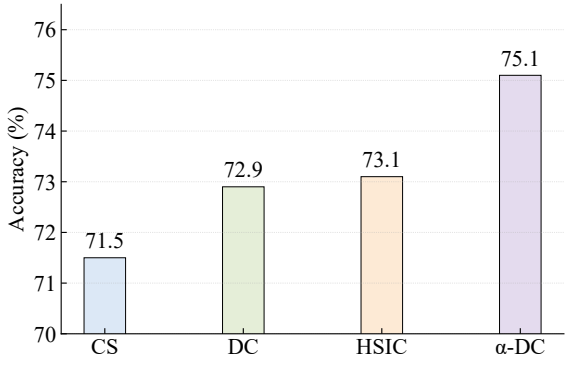

Figure 3: α-DC vs. other alternatives (1-shot on SSv2-Full)

Query-Specific	Task Prototype	SSv2-Full	HMDB51
w/o	Average	74.1	84.5
w/	Average	75.1	85.0
	Concatenation	73.6	84.6
	Cross-Attention	74.1	84.6

Table 3: Ablation on Query-Specific task prototype

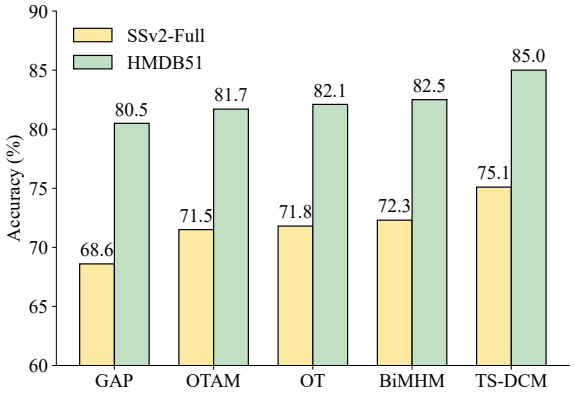

Figure 4: Comparison with different metrics

fect of different task prototypes, we conduct a comprehensive ablation study on prototype construction strategies. Specifically, we first investigate whether incorporating the query into prototype construction is beneficial. And to build query-specific prototypes, we explore various fusion strategies between support and query class tokens, including simple averaging, concatenation along the feature dimension, and cross-attention. As shown in Table 3, incorporating the query into prototype construction yields a performance gain of approximately 0.5%∼1.0% on both SSv2-Full and HMDB51. Among the query-specific strategies, simple averaging consistently outperforms concatenation and cross-attention by 0.4%∼1.5% across both datasets.

Comparison of Different Metrics To evaluate the effectiveness of the proposed TS-DCM metric, we replace it with several commonly used metrics within our framework a fair comparison. As illustrated in Figure 4, we consider Global Average Pooling (GAP), OTAM, Optimal Transport(OT), and BiMHM (Wang et al. 2022) as baselines. TS-DCM consistently achieves higher accuracy than all prior metrics on both SSv2-Full and HMDB51. In particular, compared to the second-best metric, BiMHM, TS-DCM improves performance by 2.8% on SSv2-Full and 2.5% on HMDB51. These results highlight the advantage of our metric in fully modeling inter-frame relationships and effectively leveraging task-specific information during matching.

Combine IF-D$^\alpha$C with existing metrics Our proposed IF-D$^\alpha$C enables comprehensive modeling of inter-frame relationships. To evaluate its generalization ability, we incorporate it with existing metrics and conduct the evaluation under our setting. As shown in Table 4, incorporating IF-D$^\alpha$C to model inter-frame dependencies yields performance gains of 0.8%∼3.4% over the original metrics.

Metric	SSv2-Full		HMDB51	
	w/o	w/	w/o	w/
GAP	68.6	72.0 (+3.4)	80.5	81.8 (+1.3)
OTAM	71.5	72.4 (+0.9)	81.7	83.7 (+2.0)
BiMHM	72.3	73.2 (+0.9)	82.5	84.5 (+2.0)
OT	71.8	73.5 (+1.7)	82.1	82.9 (+0.8)

Table 4: Evaluation of existing metrics with ('w/') and without ('w/o') IF-D$^\alpha$C.

Method	Time	Params	Mem	SSv2-Full	HMDB51
CLIP-FSAR	0.70 s	89 M	∼20 GB	62.1	77.1
EMP-Net	0.45 s	9 M	∼4 GB	63.1	76.8
TS-FSAR†	0.42 s	14 M	∼9 GB	75.1	85.0
TS-FSAR*	0.32 s	5 M	∼3.6 GB	67.0	84.5

Table 5: Efficiency comparison with prior methods were performed under 5-way 1-shot setting (with 2 queries). * indicates using only a 3-layer LSN, while † denotes using a 12-layer one. All evaluations were completed using a single RTX 4090 GPU.

4.5 Efficiency Analysis

To evaluate efficiency, we report the average training and inference time per task, as well as parameter count and memory usage. As shown in Table 5, with a 12-layer LSN, TS-FSAR achieves the fastest runtime (0.42 s) and markedly reduces memory (9 GB) and parameters (14 M) compared to fully fine-tuned CLIP-FSAR (0.70 s, 20 GB, 89 M), while maintaining comparable cost to EMP-Net (0.45 s, 4 GB, 9 M). When the LSN depth is reduced to 3 layers, matching EMP-Net, TS-FSAR still delivers superior performance with the fastest speed, minimal memory usage, and the fewest parameters (0.32 s, 3.6 GB, 5 M).

5 Conclusion

We propose TS-FSAR, a novel few-shot action recognition framework that introduces Task-Specific Distance Correlation Matching (TS-DCM) — a new metric designed to address key limitations of previous methods, which rely on cosine similarity to model linear inter-frame dependencies and overlook task-specific cues during matching. TS-DCM uses α-distance correlation to capture both linear and nonlinear inter-frame relationships, and employs a query-specific task prototype to enable task-specific query-support matching. To efficiently adapt CLIP, we employ a visual Ladder Side Network (LSN), whose training is guided by the adapted frozen CLIP outputs to achieve reliable correlation estimation under limited data. With these designs, TS-FSAR achieves superior performance on five standard benchmarks.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant Nos. 62471083 and 61971086).

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846.

Cao, C.; Zhang, Y.; Yu, Y.; Lv, Q.; Min, L.; and Zhang, Y. 2024. Task-Adapter: Task-specific Adaptation of Image Models for Few-shot Action Recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9038–9047.

Cao, K.; Ji, J.; Cao, Z.; Chang, C.; and Niebles, J. C. 2020. Few-Shot Video Classification via Temporal Alignment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 10615–10624.

Carreira, J.; and Zisserman, A. 2017a. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Carreira, J.; and Zisserman, A. 2017b. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 4724–4733.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Fu, Y.; Zhang, L.; Wang, J.; Fu, Y.; and Jiang, Y.-G. 2020. Depth guided adaptive meta-fusion network for few-shot video recognition. In *Proceedings of the 28th ACM international conference on multimedia*, 1142–1151.

Goyal, R.; Kahou, S. E.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fründ, I.; Yianilos, P.; Mueller-Freitag, M.; Hoppe, F.; Thurau, C.; Bax, I.; and Memisevic, R. 2017. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 5843–5851.

Gretton, A.; Bousquet, O.; Smola, A.; and Schölkopf, B. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*, 63–77. Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T. A.; and Serre, T. 2011. HMDB: A large video database for human motion recognition. In *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, 2556–2563.

Kumar, P.; Padmanabhan, N.; Luo, L.; Rambhatla, S. S.; and Shrivastava, A. 2024. Trajectory-aligned Space-time Tokens for Few-shot Action Recognition. In *European Conference on Computer Vision*, 474–493.

Lee, S.; Moon, W.; Seong, H. S.; and Heo, J.-P. 2025. Temporal Alignment-Free Video Matching for Few-shot Action Recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5412–5421.

Leyder, S.; Raymaekers, J.; and Rousseeuw, P. J. 2024. Is Distance Correlation Robust?". *arXiv preprint arXiv:2403.03722*, 122.

Li, B.; Liu, M.; Wang, G.; and Yu, Y. 2025. Frame Order Matters: A Temporal Sequence-Aware Model for Few-Shot Action Recognition. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 18218–18226.

Li, S.; Liu, H.; Qian, R.; Li, Y.; See, J.; Fei, M.; Yu, X.; and Lin, W. 2022. Ta2n: Two-stage action alignment network for few-shot action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, 1404–1411.

Lin, J.; Gan, C.; and Han, S. 2019. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7083–7093.

Lin, Z.; Geng, S.; Zhang, R.; Gao, P.; De Melo, G.; Wang, X.; Dai, J.; Qiao, Y.; and Li, H. 2022. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, 388–404.

Liu, R.; Huang, J.; Li, G.; Feng, J.; Wu, X.; and Li, T. H. 2023. Revisiting temporal modeling for clip-based image-to-video knowledge transferring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6555–6564.

Nguyen, K. D.; Tran, Q.; Nguyen, K.; Hua, B.; and Nguyen, R. 2022. Inductive and Transductive Few-Shot Video Classification via Appearance and Temporal Alignments. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XX*, 471–487. Springer.

Park, J.; Lee, J.; and Sohn, K. 2023. Dual-path adaptation from image to video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2203–2213.

Pei, W.; Tan, Q.; Lu, G.; and Tian, J. 2025. D ^{2} ST-Adapter: Disentangled-and-Deformable Spatio-Temporal Adapter for Few-shot Action Recognition. In *IEEE/CVF International Conference on Computer Vision*.

Perrett, T.; Masullo, A.; Burghardt, T.; Mirmehdi, M.; and Damen, D. 2021. Temporal-Relational CrossTransformers for Few-Shot Action Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 475–484.

Qian, R.; Ding, S.; and Lin, D. 2024. Rethinking image-to-video adaptation: An object-centric perspective. In *European Conference on Computer Vision*, 329–348. Springer.

Qu, H.; Yan, R.; Shu, X.; Gao, H.; Huang, P.; and Xie, G.-S. 2025. MVP-shot: Multi-velocity progressive-alignment framework for few-shot action recognition. *IEEE Transactions on Multimedia*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, 8748–8763.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Sung, Y.; Cho, J.; and Bansal, M. 2022. LST: Ladder Side-Tuning for Parameter and Memory Efficient Transfer Learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Székely, G. J.; and Rizzo, M. L. 2009. Brownian distance covariance. *Annals of Statistics*, 3: 1236–1265.

Thatipelli, A.; Narayan, S.; Khan, S.; Anwer, R. M.; Khan, F. S.; and Ghanem, B. 2022. Spatio-temporal Relation Modeling for Few-shot Action Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 19926–19935.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, 20–36.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Wang, X.; Zhang, S.; Cen, J.; Gao, C.; Zhang, Y.; Zhao, D.; and Sang, N. 2024. CLIP-guided Prototype Modulating for Few-shot Action Recognition. *Int. J. Comput. Vis.*, 1899–1912.

Wang, X.; Zhang, S.; Qing, Z.; Gao, C.; Zhang, Y.; Zhao, D.; and Sang, N. 2023. Molo: Motion-augmented long-short contrastive learning for few-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18011–18021.

Wang, X.; Zhang, S.; Qing, Z.; Tang, M.; Zuo, Z.; Gao, C.; Jin, R.; and Sang, N. 2022. Hybrid Relation Guided Set Matching for Few-shot Action Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 19916–19925.

Wu, C.; Wu, X.-J.; Li, L.; Xu, T.; Feng, Z.; and Kittler, J. 2024. Efficient Few-Shot Action Recognition via Multi-level Post-reasoning. In *European Conference on Computer Vision*, 38–56.

Wu, J.; Zhang, T.; Zhang, Z.; Wu, F.; and Zhang, Y. 2022. Motion-modulated Temporal Fragment Alignment Network For Few-Shot Action Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 9141–9150.

Xie, J.; Long, F.; Lv, J.; Wang, Q.; and Li, P. 2022. Joint Distribution Matters: Deep Brownian Distance Covariance for Few-Shot Classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 7962–7971.

Xing, J.; Wang, M.; Liu, Y.; and Mu, B. 2023a. Revisiting the Spatial and Temporal Modeling for Few-Shot Action Recognition. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 3001–3009.

Xing, J.; Wang, M.; Ruan, Y.; Chen, B.; Guo, Y.; Mu, B.; Dai, G.; Wang, J.; and Liu, Y. 2023b. Boosting Few-shot Action Recognition with Graph-guided Hybrid Matching. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 1740–1750.

Xing, J.; Zhao, J.; Xu, C.; Wang, M.; Dai, G.; Liu, Y.; Wang, J.; and Li, X. 2025. MA-FSAR: Multimodal Adaptation of CLIP for few-shot action recognition. *Pattern Recognition*, 111902.

Zhang, J. O.; Sax, A.; Zamir, A.; Guibas, L.; and Malik, J. 2020. Side-tuning: a baseline for network adaptation via additive side networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 698–714.

Zhelezniak, V.; Savkov, A.; Shen, A.; and Hammerla, N. Y. 2019. Correlation Coefficients and Semantic Textual Similarity. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 951–962.

Zheng, S.; Chen, S.; and Jin, Q. 2022. Few-Shot Action Recognition with Hierarchical Matching and Contrastive Learning. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, 297–313.

Zhou, B.; Andonian, A.; Oliva, A.; and Torralba, A. 2018. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, 803–818.

Zhu, L.; and Yang, Y. 2018. Compound Memory Networks for Few-Shot Video Classification. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, 782–797.