Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Active Multimodal Distillation for Few-shot Action Recognition

Weijia Feng[1], **Yichen Zhu**[1], **Ruojia Zhang**[1], **Chenyang Wang**[2,3*],
Fei Ma[3], **Xiaobao Wang**[4] and **Xiaobai Li**[5,6]
[1]College of Computer and Information Engineering, Tianjin Normal University, Tianjin, China
[2]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[3]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China
[4]College of Intelligence and Computing, Tianjin University, Tianjin, China
[5]The State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou, China
[6]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security, Hangzhou, China
weijiafeng@tjnu.edu.cn, {zhuyiichen, zrj20001127}@163.com, chenyangwang@ieee.org,
mafei@gml.ac.cn, wangxiaobao@tju.edu.cn, xiaobai.li@zju.edu.cn

Abstract

Owing to its rapid progress and broad application prospects, few-shot action recognition has attracted considerable interest. However, current methods are predominantly based on limited single-modal data, which does not fully exploit the potential of multimodal information. This paper presents a novel framework that actively identifies reliable modalities for each sample using task-specific contextual cues, thus significantly improving recognition performance. Our framework integrates an Active Sample Inference (ASI) module, which utilizes active inference to predict reliable modalities based on posterior distributions and subsequently organizes them accordingly. Unlike reinforcement learning, active inference replaces rewards with evidence-based preferences, making more stable predictions. Additionally, we introduce an active mutual distillation module that enhances the representation learning of less reliable modalities by transferring knowledge from more reliable ones. Adaptive multimodal inference is employed during the meta-test to assign higher weights to reliable modalities. Extensive experiments across multiple benchmarks demonstrate that our method significantly outperforms existing approaches.

1 Introduction

Over the past few years, video action recognition has witnessed substantial advancements, largely due to the rapid development of deep learning technologies. Although traditional approaches, including 2D and 3D Convolutional Neural Networks (CNNs), have shown remarkable proficiency in capturing both spatial and temporal features within videos, they typically demand extensive labeled data for effective training. The acquisition of such large-scale labeled datasets is often costly and time-intensive, especially for practical action recognition applications. To address this issue, recent research [Yang *et al.*, 2023] has focused on enhancing the efficiency of video understanding through lightweight models that can process temporal information efficiently while minimizing few-shot action recognition costs.

The existing methods [Feng *et al.*, 2024; Wang *et al.*, 2024a] have shown significant performance. They mainly rely on single-modal data (such as RGB frames), and human actions comprise modalities such as RGB, optical flow, skeletal nodes, and depth information. These methods cannot capture the global information of human actions, and their performance will decrease to a certain extent when encountering complex actions. Due to the limitations inherent in single-modality approaches [Ma *et al.*, 2024a], the potential of multiple modalities has attracted extensive scholarly attention and been empirically validated as effective [Xue *et al.*, 2024; Ma *et al.*, 2024b]. Several state-of-the-art methods have begun to combine multiple modalities [Wu *et al.*, 2025; Wang *et al.*, 2025; Ma *et al.*, 2024c], such as visual data, optical flow, and audio, to provide complementary recognition. However, these methods cannot identify which modality is important and which other modalities are not important in the current sample. These methods may assign extremely high weights to unimportant modalities when recognizing actions, resulting in unsatisfactory recognition results.

Inspired by Active Inference [Tschantz *et al.*, 2020], this paper proposes an Active Multimodal Few-Shot Inference for Action Recognition (AMFIR) to address the limitations of single-modal data in few-shot action recognition. The proposed AMFIR significantly improves the accuracy and efficiency of inference by actively identifying the most dominant modality of each query sample. This framework adopts a meta-learning paradigm, where each learning unit consists of labeled support samples and unlabeled query samples. In the meta-training phase, we use a modality-specific backbone network to extract feature representations based on active inference, and divide the query samples into an RGB dominant group and an optical-flow dominant group. We further designed a bidirectional distillation mechanism to guide the

*Corresponding author: Chenyang Wang.

learning of unreliable modes through reliable modes. In the meta-test phase, Active Multimodal Inference (AMI) dynamically fuses posterior distributions of different modalities, assigning higher weights to more reliable modalities to optimize inference results. Overall, the main contributions of this article can be summarized as follows:

- This article utilizes the natural complementarity between different modalities to select the most dominant modality for each query sample through active inference, thereby significantly improving the performance of few-shot action recognition.
- A mutual refinement strategy has been proposed to transfer task-related knowledge learned from reliable modes to representation learning of unreliable modalities, leveraging the complementarity of multiple modalities to enhance the ability to identify unreliable modalities.
- We have designed an adaptive multimodal few-sample inference method that combines the results of specific modalities and assigns higher weights to more reliable modalities to optimize recognition performance further.

2 Related Work

2.1 Few-shot Learning

Few-shot learning (FSL) focuses on recognizing novel concepts with minimal labeled training data. In recent years, few-shot learning (FSL) has made significant strides in various traditional domains, such as image classification [Bateni *et al.*, 2020], object detection [Fan *et al.*, 2020; Ma *et al.*, 2024c], and segmentation [Liu *et al.*, 2020; Feng *et al.*, 2025]. Despite these achievements, most existing FSL methods primarily concentrate on single-modality data, with relatively limited exploration of multimodal approaches. For instance, some techniques [Pahde *et al.*, 2019] enhance low-shot visual embeddings by incorporating auxiliary text data during training, thereby boosting performance in few-shot image classification. Others [Dong *et al.*, 2018] focus on modeling the interplay between visual and textual information to address tasks like few-shot image description and visual question answering. Similarly, [Tsimpoukelli *et al.*, 2021] leverages pre-trained language models to extend few-shot learning capabilities to downstream multimodal tasks, including visual question answering. Collectively, these studies highlight the potential of multimodal data to address the limitations inherent in unimodal FSL methods.

2.2 Few-shot Action Recognition

Two promising few-shot action recognition techniques have been proposed in the study. The first type utilizes data augmentation to support robust representation learning by creating supplementary training data [Kumar Dwivedi *et al.*, 2019], self-supervised cues [Zhang *et al.*, 2020], or auxiliary information [Fu *et al.*, 2020; Wu *et al.*, 2022]. Premier TACO [Zheng *et al.*, 2024] has improved the efficiency of the few-shot learning strategy through multi-task feature representation and negative sample selection mechanism, demonstrating significant performance. The second method focuses on alignment and strives to evaluate the similarity between query samples and support samples through synchronization frames or periods in the temporal or spatial domain [Cao *et al.*, 2020; Wang *et al.*, 2022; Wu *et al.*, 2022]. The M2-CLIP framework [Wang *et al.*, 2024b] utilizes multimodal adapters and multitasking decoders to improve video action recognition while maintaining a strong zero sample generalization ability.

2.3 Active Inference

Active Inference (AIF) is a Bayesian framework explaining how organisms minimize uncertainty and surprise by predicting and evaluating sensory inputs. It originates from Karl Friston's Free Energy Principle (FEP) and emphasizes how organisms predict future states and adjust behavior to achieve goals through generative models during environmental interactions. For example, [Sedlak *et al.*, 2024] explored the application of active reasoning in edge computing, demonstrating its potential in adaptive flow processing to meet service level objectives (SLOs) and real-time system management. Similarly, [Pezzulo *et al.*, 2023] highlighted the close relationship between active reasoning and Generative AI, emphasizing its role in achieving higher-level intelligence and understanding through active perception and action. These advancements suggest that active reasoning provides a novel perspective for designing and implementing artificial intelligence, fostering progress in related fields. Furthermore, in hyperspectral image classification (HSIC), the Active Inference Transfer Convolutional Fusion Network (AI-TFNet) proposed by [Wang *et al.*, 2023] utilizes a pseudo-label propagation algorithm to enhance the availability of training samples and optimize classification performance.

2.4 Knowledge Distillation

Knowledge distillation serves as an effective knowledge transfer technique, extracting information from teacher networks and conveying it to student models. Recent research [Hinton, 2015; Park *et al.*, 2019; Tung and Mori, 2019] has highlighted its potential in cross-modal tasks. For instance, [Gupta *et al.*, 2016] developed a method to transfer supervision across different modalities, using representations learned from well-labeled modalities as guidance for training new, unlabeled modalities. Similarly, [Garcia *et al.*, 2018] proposed a technique for training multimodal video action recognition models using both depth and RGB data, addressing challenges such as noise and missing modalities during tests. The MARS framework [Crasto *et al.*, 2019] simulates optical flow through RGB frame training, avoiding optical flow computation at test time. It integrates appearance and motion information by combining feature loss and cross-entropy loss. Additionally, [Dai *et al.*, 2021] introduced a knowledge distillation framework for action detection, enhancing RGB representations by leveraging knowledge from other modalities like optical flow and 3D pose. This approach achieves performance comparable to dual-stream networks using only RGB data during the test.

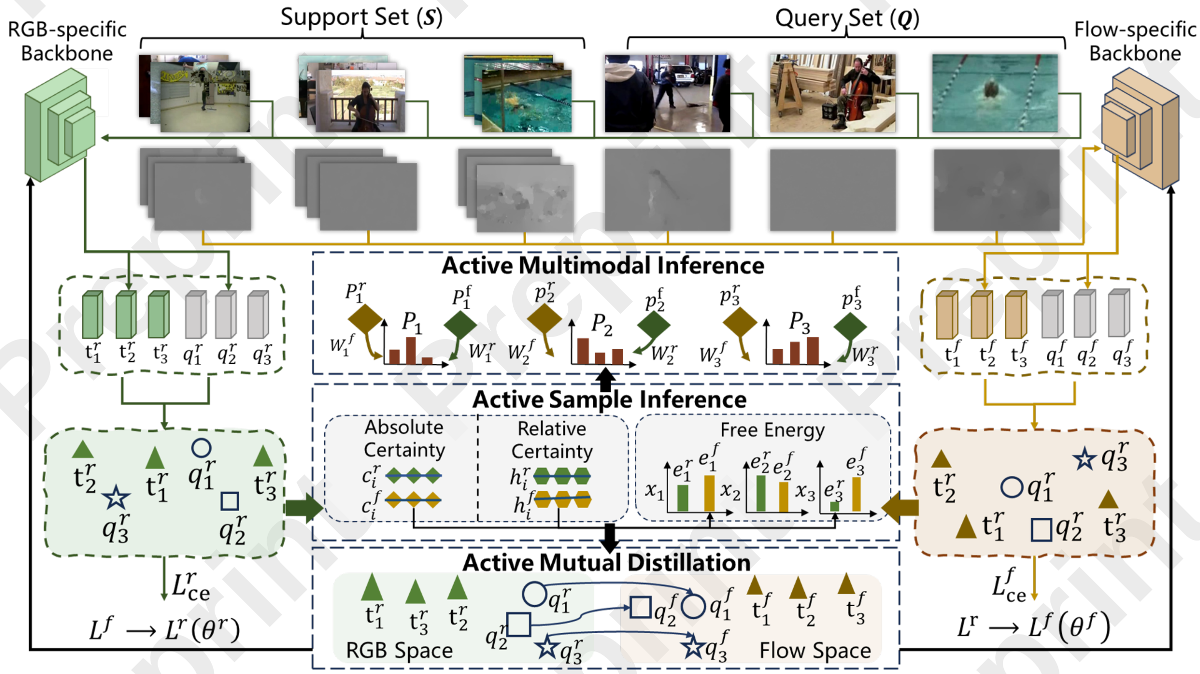


Figure 1: Illustration of our proposed framework in the 3-way 3-shot setting.

3 Method

3.1 Problem Definition

This paper employs a meta-learning framework for few-shot multimodal action recognition, which includes two primary phases: meta-training and meta-test. In the meta-training phase, we utilize a multimodal video dataset D_{train} that encompasses base action classes C_{train}. We construct multiple meta-tasks (often termed episodes) from D_{train} to train a meta-learner capable of generalizing to new action classes. Each meta-task τ consists of a query set $Q \subset D_{train}$ and a support set $S \subset D_{train}$. Within the $N-way\ K-shot$ meta-learning setting, the query set $Q = \{(x_i^r, x_i^f, y_i)\}_{i=1}^{M}$ includes M multimodal query samples. Here, x_i^r and x_i^f represent two modalities (RGB and optical flow) of the $i-th$ query sample, and $y_i \in \{1, 2, \cdots, N\}$ denotes the class label of the $i-th$ query sample. The support set $S = \{(x_i^r, x_i^f, y_i)\}_{i=M+1}^{M+NK}$ contains K multimodal samples for each of the N classes. In the meta-test phase, we employ a multimodal dataset D_{test}, which includes novel action classes C_{test} that are disjoint from the training classes ($C_{test} \cap C_{train} = \oslash$). Similar to the meta-training phase, the support and query sets for each test task are constructed in the same manner. A key point to highlight is that the class labels of the query samples are concealed during the meta-test. The meta-learner must accurately classify each sample in the query set, relying exclusively on the labeled samples provided in the support set.

3.2 Overview

The overall architecture of the proposed AFMIR is depicted in Figure 1. For each episode, we utilize a backbone network $\phi^m(Q, S; \theta^m), m \in (r, f)$ to extract the feature representations of query samples $\{q_i^m\}_{i=1}^{M}$ and the prototypes of support samples $\{t_i^m\}_{k=1}^{N}$ for each modality. Subsequently, we calculate the modality-specific posterior distributions for each query sample based on the distances between query samples and prototypes in the modality-specific feature space. During the meta-training phase, we introduce an Active Sample Inference (ASI) module, which takes the modality-specific posterior distributions as inputs and performs reliability inference to assess the reliability of each modality (RGB and optical flow) for each query sample. This process categorizes the samples into two groups: the optical flow dominant group $\mathcal{G}^f$ and RGB dominant group $\mathcal{G}^r$. For the selected samples in $\mathcal{G}^f$ and $\mathcal{G}^r$, we implement Active Mutual Distillation to transfer task-specific knowledge from reliable modalities to unreliable ones via a bidirectional distillation mechanism, thereby enhancing the representation learning of the less dominant modality. In the meta-test phase, we employ adaptive multimodal inference, which leverages modality-specific posterior distributions to make adaptive fusion decisions, prioritizing the more reliable modalities.

3.3 Active Sample Inference

In this module, we use the idea of AIF to actively predict the most dominant modality of each sample, and this predicted modality will be considered as the sample-specific dominant modality of the corresponding sample. Before introducing this module, we first need to review the method of AIF, which aims to maximize the Bayesian model evidence for an agent's generative model in the context of Partially Observed Markov Decision Processes (POMDP). Formally, a POMDP is defined by a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{P}, \Theta \rangle$, where $\mathcal{S}$ denotes the true state of the environment and $\mathcal{A}$ denote agent's action space.

During each time step t, the agent transitions to a new state s_t, which is calculated through $P(s_t|s_{t-1}, a_{t-1})$, where the transition is informed by $s_{t-1} \in \mathcal{S}$ and $a_{t-1} \in \mathcal{A}$ executed at the preceding time step $t-1$. Agents might not have direct access to the actual state of the environment. Instead, they receive observations, denoted as o_t, sampled from a distribution $P(o_t \mid s_t)$ that depends on the true environmental state s_t. Given this limitation, agents must base their operations on an inferred belief about the true state, $\hat{s}_t$, representing their estimation of s_t based on the received observations. Within the framework of a specified generative model, agents perform approximate Bayesian inference through the encoding of an arbitrary probability distribution $q(s, \theta)$, which is optimized through the minimization of variational free energy $\tilde{\mathcal{F}}$:

$$\tilde{\mathcal{F}} = D_{KL}\left(q(s, \theta) \| p^{\Phi}(o, s, \theta)\right), \tag{1}$$

where $o \in \mathcal{O}$ denotes the agent's observations and $\theta \in \Theta$ represents model parameters.

To establish the active inference model, we input the representations of query samples $\{q_i^m\}_{i=1}^{M}$ and the prototypes of support samples $\{t_k^m\}_{k=1}^{N}$ into the module. We consider the dominant modality for each query sample as the one that can reflect more task-specific discriminative features. However, the dominant modality for each query sample is not fixed, as the contribution of a specific modality largely depends on the contextual information of the query and support samples in each task. To address this, we propose to infer the reliability of different modalities based on modality-specific posterior distributions. This approach aligns with the core principles of active inference, where the model dynamically assesses the reliability of different modalities to optimize its predictive capabilities. For each query sample, the modality-specific posterior distribution p_i^m can be formulated as:

$$p_i^m(k|x_i^m) = \frac{e^{-\psi(q_i^m, t_k^m)}}{\sum_{k'=1}^{N} e^{-\psi(q_i^m, t_{k'}^m)}}, \tag{2}$$

where $k \in \{1, \ldots, N\}, m \in \{r, f\}$, and q_i^m denotes the modality-specific representation of the i-th query sample, t_k^m denotes the prototype of the k-th class, and ψ is a distance measurement function (*e.g.,* Euclidean distance in this work). The posterior distribution reflects the model's belief in the class labels given the specific modality, consistent with minimizing prediction errors to optimize beliefs. By dynamically evaluating the reliability of each modality in the context of the current task, the model can select the most informative modality to reduce prediction errors.

After obtaining the modal-specific posterior distribution p_i^m for the i^{th} query sample, we use it as input to construct the observation space $\mathcal{O}$ and the state space $\mathcal{S}$. The reliability of each modality for a given sample can be assessed by the free energy derived from Eq. (1). To elaborate, the modality with the lower free energy value would be considered more reliable according to the calculations based on the provided formula. This approach enables a quantitative comparison between different modalities, facilitating a more accurate determination of their respective reliabilities. In order to systematically investigate the cross-modal complementarity, a selection criterion is established based on the significant discrepancy in reliability between the two modalities. Consequently, the queried samples are meticulously categorized into the following distinct groups:

$$\mathcal{G}^m = \{(x_i^r, x_i^f) \mid (x_i^r, x_i^f) \in \mathcal{G},\ \mathcal{F}_i^m > \mathcal{F}_i^n\}, \tag{3}$$

where $m, n \in \{r, f\},\ m \neq n$, and $\mathcal{G}^m$ denotes the group dominated by modality m, comprising query samples for which the modality m exhibits higher certainty than modality n in the few-shot task. Specifically, when $m = f$, $\mathcal{G}^m$ represents the Flow-dominant group. Conversely, when $m = r$, $\mathcal{G}^m$ corresponds to the RGB-dominant group, indicating that the RGB modality is more reliable for distinguishing between query samples.

3.4 Active Mutual Distillation

In this section, we first introduce an active mutual distillation method, which enhances the representation learning of less reliable modes by utilizing task-specific discriminative knowledge from more reliable modes. Traditional knowledge extraction methods typically involve using well-trained teacher models to guide student model learning through consistency constraints, such as KL divergence calculated based on logits:

$$D_{KL}(p^t \| p^s) = \sum_{i=1}^{N} p_i^t \left(\log p_i^t - \log p_i^s\right), \tag{4}$$

where p^t and p^s represent the logits produced by the teacher and student models, respectively. In conventional methods, teacher-student distillation is consistently applied to individual samples.

Based on this, we define the absolute certainty c_i^m as the maximum value of the modality-specific posterior distribution, which quantifies the model's highest confidence in any class prediction for modality m:

$$c_i^m = \max_k p_i^m(k|x_i^m), \tag{5}$$

where $p_i^m(k|x_i^m)$ denotes the posterior distribution over classes for the i-th query sample in modality m. Absolute certainty c_i^m reflects the highest belief of the model in its prediction for that modality. This dynamic assessment allows us to adaptively evaluate the reliability of different modalities based on contextual information in each task, thus better accommodating the specific demands of different tasks.

To exploit the complementarity between different modalities and improve few-shot action recognition, we refine the approach by treating models trained on one modality as teachers and those trained on another modality as students. However, determining which modality should act as the teacher is challenging, as the contribution of each modality varies between samples and depends heavily on the contextual information of the few-shot task. To address this, we previously introduced active inference to dynamically infer the importance of different modalities for each sample. We actively assign more reliable modalities as teachers to transfer knowledge. Specifically, we constrain the learning of two modality-specific models by actively transferring query-to-prototype relationship knowledge between different modalities:

$$\mathcal{L}_{m\to n}(\theta^n) = \frac{1}{\sum_{(x_i^n, x_i^m)} c_i^m} \sum_{(x_i^n, x_i^m)} c_i^m D_{KL}(p_i^m, p_i^n), \quad (6)$$

where p_i^m denotes the modality-specific posterior distribution for the i-th query sample in modality m, as defined in Eq.(2), c_i^m represents the absolute certainty for the i-th query sample in modality m, as defined in Eq. (5), and $m, n \in \{r, f\}$ with $m \neq n$.

3.5 Active Multimodal Inference

In the first two sub-sections, we explored how different modalities of each sample contribute to meta-training and discussed their role in the meta-test phase. Specifically, we introduce a method for adaptively integrating multimodal predictions to form the final decision in few-shot inference. Given the varying reliability of modalities for each query sample, we design an adaptive multimodal fusion strategy as follows:

$$\mathcal{P}(k|x_i^m) = \frac{e^{-\alpha_i^r \psi(q_i^r, t_k^r) - \alpha_i^f \psi(q_i^f, t_k^f)}}{\sum_{k'=1}^{N} e^{-\alpha_i^r \psi(q_i^r, t_{k'}^r) - \alpha_i^f \psi(q_i^f, t_{k'}^f)}}, \quad (7)$$

where $\psi(\cdot)$ represents the Euclidean distance function, and α_i^r and α_i^f are the adaptive fusion weights for the RGB and optical flow modalities of the i-th query sample, respectively.

Since modality-specific posterior distributions may not always be accurate during meta-tests, and relative certainty does not directly reflect the similarity between query samples and class prototypes, we use the absolute certainty values c_i^m to compute the adaptive fusion weights:

$$\alpha_i^m = \frac{c_i^m}{c_i^r + c_i^f}, \quad (8)$$

where $m \in \{r, f\}$ indicates the modality, where r represents the RGB modality and f represents the optical flow modality.

3.6 Optimization

The proposed AFMIR can be optimized with the following function:

$$\mathcal{L} = \sum_{m\in\{r,f\}} \mathcal{L}_{ce}^m(\theta^m) + \lambda \sum_{m\neq n\in\{r,f\}} \mathcal{L}^m \to \mathcal{L}^n(\theta^n), \quad (9)$$

where $\mathcal{L}_{ce}^m(\theta^m)$ denotes the cross-entropy loss for modality m, and $\mathcal{L}^m \to \mathcal{L}^n(\theta^n)$ represents the knowledge distillation loss from modality m to modality n. The parameter λ balances the contribution of the distillation losses. Besides, the $\mathcal{L}_{ce}^m(\theta^m)$ is further used to constrain the modality-specific predictions, *i.e.,*

$$\mathcal{L}_{ce}^m(\theta^m) = \sum_{i=1}^{M}\sum_{k=1}^{N} p_i^m(k) \log p_i^m(k), \quad (10)$$

where $m \in \{r, f\}$ indicates the modality, r represents the RGB modality and f represents the optical flow modality. The cross-entropy loss $\mathcal{L}_{ce}^m(\theta^m)$ is computed for each modality based on the predicted probabilities $p_i^m(k)$ of the query samples belonging to each class k.

The parameters of the modality-specific backbone networks $\theta^m, m \in \{r, f\}$ are optimized via distinct weighted combinations of their corresponding losses:

$$\theta^m \leftarrow \theta^m - \gamma \nabla_{\theta^m} (\mathcal{L}_{ce}^m(\theta^m) + \lambda \sum_{n\neq m\in\{r,f\}} \mathcal{L}^n \to \mathcal{L}^m(\theta^m)). \quad (11)$$

The parameter update for each modality θ^m is performed by minimizing the cross-entropy loss $\mathcal{L}_{ce}^m(\theta^m)$ and the knowledge distillation loss from the other modality $n \neq m$. The learning rate is denoted by γ, and λ balances the contribution of the distillation losses.

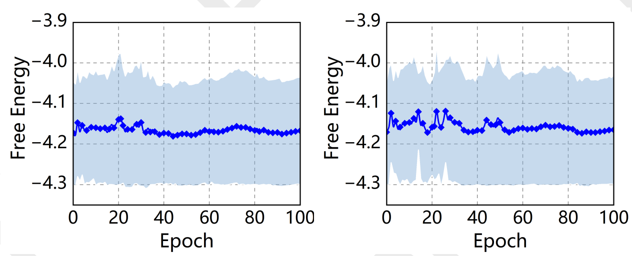
4 Experiments

4.1 Validation Protocol

Datasets. We assess our proposed method on four prominent and challenging benchmarks for few-shot action recognition: Kinetics-400 [Kay *et al.*, 2017], Something-Something V2 [Goyal *et al.*, 2017], HMDB51 [Wang *et al.*, 2015], and UCF101 [Peng *et al.*, 2018]. These datasets are augmented to include multi-modal data by generating optical flow frame sequences from the original videos using a dense optical flow algorithm.

Modality-specific Backbones. For the RGB modality, a pre-trained ResNet-50 backbone is employed to extract visual features at the frame level. For the optical flow modality, an I3D model pre-trained on the Charades dataset is used to capture motion features from individual frames. Subsequently, video-level features for both modalities are derived by aggregating these enhanced frame-level features. Ultimately, the query-specific prototype is generated by consolidating the video-level features from the support samples of the corresponding action class.

Parameters. In the meta-training phase, the balance weight (λ, specified in Eq. 9) is uniformly set to 1.0 across all benchmarks. Training is conducted using the SGD optimizer. For both RGB and optical flow modalities, the respective networks are iteratively updated by minimizing a combined weighted loss function, which includes both cross-entropy and distillation losses, until convergence is achieved. The learning rate γ is set as 10^{-3}.



(a) SSv2 (b) Kinetics-400

Figure 2: The change of free energy.

4.2 Free Energy

The experimental results show that, on the SSv2 and Kinetics-400 datasets, there is a significant fluctuation in the initial free

Modality	Method	Kinetics		SSv2		HMDB51		UCF101	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
RGB	Matching Net	53.3	78.9	-	-	-	-	-	-
	TRX	63.6	85.9	42.0	64.6	-	75.6	-	96.1
	HyRSM	73.7	86.1	54.3	69.0	60.3	78.0	83.9	94.7
	STRM	-	86.7	-	68.1	-	77.3	-	96.9
Flow	ProtoNet-F	45.2	69.5	32.9	51.1	43.7	65.0	69.7	89.6
	TRX-F	44.8	69.7	30.7	52.4	43.0	67.6	65.6	90.6
	STRM-F	47.8	69.7	36.3	55.7	52.2	67.9	79.7	91.6
Multimodal	ProtoNet-EC	63.8	84.1	33.0	49.5	56.9	73.8	78.3	93.9
	ProtoNet-EA	61.7	83.9	31.1	50.5	53.2	46.3	76.7	94.3
	STRM-EC	68.3	87.4	45.5	66.7	59.3	78.3	87.4	96.3
	AFMAR	80.1	92.6	61.7	79.5	73.9	87.8	91.2	99.0
	AFMIR(ours)	**82.8**	**96.1**	**70.6**	**92.3**	**83.7**	**90.0**	**94.9**	**99.1**

Table 1. Comparison with state-of-the-art few-shot action recognition methods. The best results are in bold. For multi-modal methods extended from existing unimodal methods, "EC" represents cascaded early fusion schemes, "EA" represents collaborative attention early fusion. The '-' indicates that the result is not available in published works.

energy during training. However, as training progresses, the free energy gradually decreases and tends to stabilize, eventually converging to around -4.0 on both datasets (as shown in Figure 2). This trend indicates that the model dynamically evaluates modal reliability through active inference modules, prioritizing the selection of modes with lower free energy for inference, which significantly optimizes the modal selection strategy. In addition, the knowledge distillation and dynamic weighting mechanisms between multiple modalities further reduce the uncertainty of modal selection, enhance task adaptation capability, and improve the stability of cross-modal inference. These results validate the crucial role of active inference and multimodal interaction in few-shot action recognition, enabling the model to more efficiently and reliably utilize modal information to complete tasks.

4.3 Comparative Experiments

We selected four RGB-based algorithms: Matching Net [Zhu and Yang, 2018], TRX [Perrett *et al.*, 2021], HyRSM [Wang *et al.*, 2022], and STRM [Thatipelli *et al.*, 2022]. These methods leverage metric learning, temporal relationships, spatiotemporal modeling, and hybrid relationships to advance few-shot action recognition. We retrained vision-based methods like TRX-F and STRM-F for optical flow data due to the lack of existing methods. We extended visual methods for multimodal baselines via early fusion (EC, EA) and late fusion (LF). Additionally, the chosen AFMAR [Wanyan *et al.*, 2023] algorithm enhances few-shot recognition by actively selecting reliable modalities, distilling knowledge bidirectionally, and adaptively fusing multimodal data.

As shown in Table 1, our AMFIR framework significantly outperforms existing methods in Few Shot Action Recognition tasks on datasets such as SSv2, HMDB51, UCF101, and Kinetics-400. The accuracy of 1-shot and 5-shot tasks reaches 70.6% and 92.3% (SSv2), 83.7% and 90.0% (HMDB51), 94.9% and 99.1% (UCF101), and 82.8% and 96.1% (Kinetics-400), respectively. This outstanding performance is attributed to the active sample inference module (ASI) in the framework, which dynamically selects the most reliable mode to reduce uncertainty, the active mutual distillation module (AMD) improves the representation ability of unreliable modes through bidirectional knowledge distillation, and the Active Multimodal Inference module (AMI) optimizing the complementarity of modes through adaptive fusion, thus fully utilizing the potential of multimodal data.

ASI		AMD	AMI	Kinetics		SSv2	
RD	FD			1-shot	5-shot	1-shot	5-shot
×	✓	✓	✓	64.10	68.94	58.03	66.81
✓	×	✓	✓	59.63	75.27	62.91	66.06
✓	✓	×	✓	71.64	89.05	61.81	83.37
✓	✓	✓	×	59.34	65.78	62.59	85.33
✓	✓	✓	✓	**82.85**	**96.11**	**70.59**	**92.32**

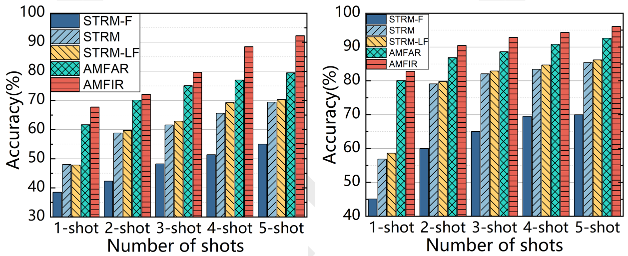
Table 2. Comparison of results across different configurations. The best are in bold.

4.4 Ablation Study

The impact of key components. The experimental results in Table 2 have been validated through ablation studies, demonstrating that the complete framework has achieved state-of-the-art performance on both Kinetics and SSv2. When the ASI module uses only RGB as the dominant modality (RD) or only optical flow as the dominant modality, the accuracy will significantly decrease. Removing AMD will reduce the thermal accuracy of Kinetics 5 by 7.06%. Compared with dynamics, disabling AMI resulted in a greater decrease in SSv2, indicating the necessity of time-sensitive adaptive fusion.

4.5 Further Remarks

Performance with Different Numbers of Support Samples. The experimental results in Figure 3 show that the performance of AMFIR steadily improves as the number of support samples changes from 1-shot to 5-shot. The accuracy on the SSv2 and Kinetics datasets increases from about 70% and 82% to about 92% and 96%, respectively, which is significantly better than other methods, especially under few-shot conditions. The reason for these results is that the ASI module reduces uncertainty by dynamically selecting reliable modalities, enabling the model to have high initial performance at 1-shot. The AMD module utilizes bidirectional knowledge distillation to enhance the collaborative effect between modalities, supporting further performance improvement as the sample size increases. The AMI module adaptively adjusts modal weights and optimizes the integration strategy of multimodal data. The synergistic effect of these modules fully utilizes the complementarity of multimodal data, resulting in the excellent performance of AMFIR in both few-shot and multi-sample scenarios.

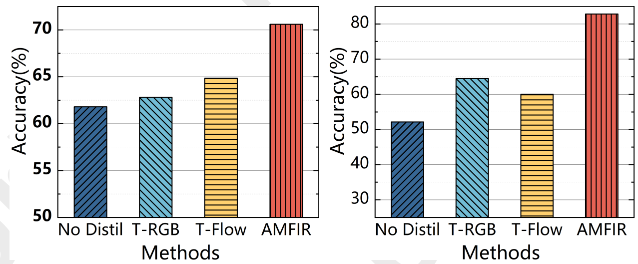


(a) SSv2 (b) Kinetics

Figure 3: Comparison results with different numbers of support samples in 5-way K-shot setting.

Influence of Different Distillation Strategies. The experimental results in Figure 4 show that the bidirectional distillation strategy (AMFIR) achieved accuracies of 70.6% and 82.8% in the 5-way 1-shot task on the SSv2 and Kinetics-400 datasets, respectively, significantly outperforming the unidirectional distillation strategy (T-RGB and T-Flow) and the no-distillation strategy (No Distillation). The reason is that bidirectional distillation effectively transmits task-related knowledge of reliable modalities through the mechanism of mutual teaching, enhances the representation ability of unreliable modalities, and further optimizes the learning effect of reliable modalities. In addition, this strategy fully utilizes the complementarity between RGB and optical flow modalities and combines active inference to dynamically evaluate the modal reliability, flexibly optimizing the distillation process. This design demonstrates the significant advantages of bidirectional distillation in few-shot multimodal learning.

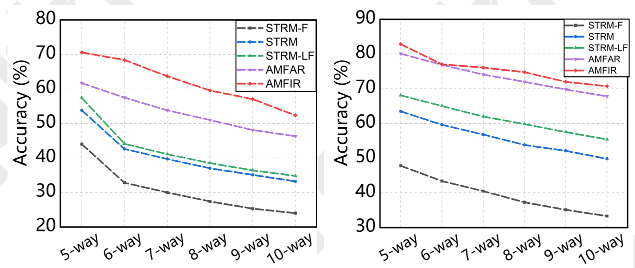
N-way Few-Shot Classification. In the N-way Few-Shot Classification task, the AMFIR framework significantly outperforms existing methods such as STRM, STRM-F, STRM-LF, and AMFAR on the SSv2 and Kinetics datasets, achieving the best accuracy in 5-way to 10-way classification tasks. The experimental results are shown in Figure 5. In the 5-way task of the SSv2 dataset, AMFIR achieved an accuracy of around 70%, while other methods were below 60%. In the Kinetics dataset, AMFIR still achieved an accuracy of around 70% in the 10-way task, significantly ahead of other methods. Its advantages are mainly due to the complementarity between RGB and optical flow modalities, the dynamic adaptive adjustment of the active inference module, and the enhancement of the representation of unreliable modalities by the bidirectional distillation strategy. This enables the framework to still



(a) SSv2 (b) Kinetics-400

Figure 4: Comparison with conventional distillation strategies in a 5-way 1-shot setting. T-RGB (or T-Flow) denotes distillation, whereas RGB (optical flow) is consistently regarded as the teacher.

exhibit strong robustness and generalization, even when task complexity increases, verifying its excellent performance in small-sample multi-classification tasks.



(a) SSv2 (b) Kinetics

Figure 5: N-way 1-shot performance on SSv2 and Kinetics.

5 Conclusion

We have proposed the Active Multimodal Few-Shot Inference for Action Recognition (AMFIR) framework that significantly enhances few-shot action recognition by actively identifying and utilizing the most reliable modalities for each sample. By integrating active mutual distillation and adaptive multimodal inference, AMFIR effectively improves the representation learning of unreliable modalities and outperforms existing methods across multiple benchmarks. This framework highlights the potential of active inference and knowledge distillation in advancing multimodal few-shot learning through uncertainty-driven modality selection, bidirectional knowledge transfer, and context-aware fusion. Extensive experiments validate its robustness in handling sensor noise, motion ambiguity, and extreme data scarcity while maintaining computational efficiency.

Acknowledgements

This study is funded in part by the National Key Research and Development Plan: 2019YFB2101900, NSFC (Natural Science Foundation of China): 61602345, 62002263, 62302333, TianKai Higher Education Innovation Park Enterprise R&D Special Project: 23YFZXYC00046, and 2024 China University Industry-Academia-Research Innovation Fund: 2024HY015.

References

[Bateni *et al.*, 2020] Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. Improved few-shot visual classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14493–14502, 2020.

[Cao *et al.*, 2020] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10618–10627, 2020.

[Crasto *et al.*, 2019] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7882–7891, 2019.

[Dai *et al.*, 2021] Rui Dai, Srijan Das, and François Bremond. Learning an augmented rgb representation with cross-modal knowledge distillation for action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13053–13064, 2021.

[Dong *et al.*, 2018] Xuanyi Dong, Linchao Zhu, De Zhang, Yi Yang, and Fei Wu. Fast parameter adaptation for few-shot image captioning and visual question answering. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 54–62, 2018.

[Fan *et al.*, 2020] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4013–4022, 2020.

[Feng *et al.*, 2024] Weijia Feng, Ruojia Zhang, Yichen Zhu, Chenyang Wang, Chuan Sun, Xiaoqiang Zhu, Xiang Li, and Tarik Taleb. Exploring collaborative diffusion model inferring for aigc-enabled edge services. *IEEE Transactions on Cognitive Communications and Networking*, 2024.

[Feng *et al.*, 2025] Weijia Feng, Xinyu Zuo, Ruojia Zhang, Yichen Zhu, Chenyang Wang, Jia Guo, and Chuan Sun. Federated deep reinforcement learning for multimodal content caching in edge-cloud networks. *IEEE Transactions on Network Science and Engineering*, 2025.

[Fu *et al.*, 2020] Yuqian Fu, Li Zhang, Junke Wang, Yanwei Fu, and Yu-Gang Jiang. Depth guided adaptive meta-fusion network for few-shot video recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1142–1151, 2020.

[Garcia *et al.*, 2018] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018.

[Goyal *et al.*, 2017] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[Gupta *et al.*, 2016] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016.

[Hinton, 2015] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Kay *et al.*, 2017] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[Kumar Dwivedi *et al.*, 2019] Sai Kumar Dwivedi, Vikram Gupta, Rahul Mitra, Shuaib Ahmed, and Arjun Jain. Protogan: Towards few shot learning for action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[Liu *et al.*, 2020] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. Crnet: Cross-reference networks for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4165–4173, 2020.

[Ma *et al.*, 2024a] Fei Ma, Yukan Li, Yifan Xie, Ying He, Yi Zhang, Hongwei Ren, Zhou Liu, Wei Yao, Fuji Ren, Fei Richard Yu, et al. A review of human emotion synthesis based on generative technology. *arXiv preprint arXiv:2412.07116*, 2024.

[Ma *et al.*, 2024b] Fei Ma, Yucheng Yuan, Yifan Xie, Hongwei Ren, Ivan Liu, Ying He, Fuji Ren, Fei Richard Yu, and Shiguang Ni. Generative technology for human emotion recognition: A scoping review. *Information Fusion*, page 102753, 2024.

[Ma *et al.*, 2024c] Huiying Ma, Dongxiao He, Xiaobao Wang, Di Jin, Meng Ge, and Longbiao Wang. Multi-modal sarcasm detection based on dual generative processes. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 2279–2287, 2024.

[Pahde *et al.*, 2019] Frederik Pahde, Oleksiy Ostapenko, Patrick Jä Hnichen, Tassilo Klein, and Moin Nabi. Self-paced adversarial training for multimodal few-shot learning. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 218–226. IEEE, 2019.

[Park *et al.*, 2019] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019.

[Peng *et al.*, 2018] Yuxin Peng, Yunzhen Zhao, and Junchao Zhang. Two-stream collaborative learning with spatial-temporal attention for video classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(3):773–786, 2018.

[Perrett *et al.*, 2021] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 475–484, 2021.

[Pezzulo *et al.*, 2023] Giovanni Pezzulo, Thomas Parr, Paul Cisek, Andy Clark, and Karl Friston. Generating meaning: Active inference and generative ai. *PsyArXiv. June*, 8, 2023.

[Sedlak *et al.*, 2024] Boris Sedlak, Victor Casamayor Pujol, Andrea Morichetta, Praveen Kumar Donta, and Schahram Dustdar. Adaptive stream processing on edge devices through active inference. *arXiv preprint arXiv:2409.17937*, 2024.

[Thatipelli *et al.*, 2022] Anirudh Thatipelli, Sanath Narayan, Salman Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Bernard Ghanem. Spatio-temporal relation modeling for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19958–19967, 2022.

[Tschantz *et al.*, 2020] Alexander Tschantz, Beren Millidge, Anil K Seth, and Christopher L Buckley. Reinforcement learning through active inference. *arXiv preprint arXiv:2002.12636*, 2020.

[Tsimpoukelli *et al.*, 2021] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

[Tung and Mori, 2019] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1365–1374, 2019.

[Wang *et al.*, 2015] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4305–4314, 2015.

[Wang *et al.*, 2022] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. Hybrid relation guided set matching for few-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19948–19957, 2022.

[Wang *et al.*, 2023] Jianing Wang, Linhao Li, Yichen Liu, Jinyu Hu, Xiao Xiao, and Bo Liu. Ai-tfnet: Active inference transfer convolutional fusion network for hyperspectral image classification. *Remote Sensing*, 15(5):1292, 2023.

[Wang *et al.*, 2024a] Chenyang Wang, Hao Yu, Xiuhua Li, Fei Ma, Xiaofei Wang, Tarik Taleb, and Victor CM Leung. Dependency-aware microservice deployment for edge computing: A deep reinforcement learning approach with network representation. *IEEE Transactions on Mobile Computing*, 2024.

[Wang *et al.*, 2024b] Mengmeng Wang, Jiazheng Xing, Boyuan Jiang, Jun Chen, Jianbiao Mei, Xingxing Zuo, Guang Dai, Jingdong Wang, and Yong Liu. A multimodal, multi-task adapting framework for video action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5517–5525, 2024.

[Wang *et al.*, 2025] Xiaobao Wang, Yujing Wang, Dongxiao He, Zhe Yu, Yawen Li, Longbiao Wang, Jianwu Dang, and Di Jin. Elevating knowledge-enhanced entity and relationship understanding for sarcasm detection. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[Wanyan *et al.*, 2023] Yuyang Wanyan, Xiaoshan Yang, Chaofan Chen, and Changsheng Xu. Active exploration of multimodal complementarity for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6492–6502, 2023.

[Wu *et al.*, 2022] Jiamin Wu, Tianzhu Zhang, Zhe Zhang, Feng Wu, and Yongdong Zhang. Motion-modulated temporal fragment alignment network for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9151–9160, 2022.

[Wu *et al.*, 2025] Sheng Wu, Dongxiao He, Xiaobao Wang, Longbiao Wang, and Jianwu Dang. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1601–1609, 2025.

[Xue *et al.*, 2024] Haiwei Xue, Xiangyang Luo, Zhanghao Hu, Xin Zhang, Xunzhi Xiang, Yuqin Dai, Jianzhuang Liu, Zhensong Zhang, Minglei Li, Jian Yang, et al. Human motion video generation: A survey. *Authorea Preprints*, 2024.

[Yang *et al.*, 2023] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*, 2023.

[Zhang *et al.*, 2020] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip HS Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 525–542. Springer, 2020.

[Zheng *et al.*, 2024] Ruijie Zheng, Yongyuan Liang, Xiyao Wang, Shuang Ma, Hal Daumé III, Huazhe Xu, John Langford, Praveen Palanisamy, Kalyan Shankar Basu, and Furong Huang. Premier-taco is a few-shot policy learner: Pretraining multitask representation via temporal action-driven contrastive loss. In *Forty-first International Conference on Machine Learning*, 2024.

[Zhu and Yang, 2018] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 751–766, 2018.