This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Beyond Label Semantics: Language-Guided Action Anatomy for Few-shot Action Recognition

Zefeng Qian[1] Xincheng Yao[1] Yifei Huang[3,4] Chongyang Zhang[1,2] * Jiangyong Ying[5] Hong Sun[5]
[1]School of Information Science and Electronic Engineering, Shanghai Jiao Tong University
[2]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[3]The University of Tokyo [4]Shanghai AI Laboratory [5]E-surfing Vision Technology Co., Ltd
{zefeng_qian, i-Dover, sunny_zhang}@sjtu.edu.cn, hyf@iis.u-tokyo.ac.jp
{yingjiangyong, sunhong}@chinatelecom.cn

Abstract

*Few-shot action recognition (FSAR) aims to classify human actions in videos with only a small number of labeled samples per category. The scarcity of training data has driven recent efforts to incorporate additional modalities, particularly text. However, the subtle variations in human posture, motion dynamics, and the object interactions that occur during different phases, are critical inherent knowledge of actions that cannot be fully exploited by action labels alone. In this work, we propose **L**anguage-**G**uided Action **A**natomy (LGA), a novel framework that goes beyond label semantics by leveraging Large Language Models (LLMs) to dissect the essential representational characteristics hidden beneath action labels. Guided by the prior knowledge encoded in LLM, LGA effectively captures rich spatiotemporal cues in few-shot scenarios. Specifically, for text, we prompt an off-the-shelf LLM to anatomize labels into sequences of atomic action descriptions, focusing on the three core elements of action (subject, motion, object). For videos, a Visual Anatomy Module segments actions into atomic video phases to capture the sequential structure of actions. A fine-grained fusion strategy then integrates textual and visual features at the atomic level, resulting in more generalizable prototypes. Finally, we introduce a Multimodal Matching mechanism, comprising both video-video and video-text matching, to ensure robust few-shot classification. Experimental results demonstrate that LGA achieves state-of-the-art performance across multiple FSAR benchmarks.*

1. Introduction

Few-shot action recognition seeks to classify human actions in videos using only a limited number of labeled samples, addressing the challenge of data scarcity in real-world applications. Building on advances in few-shot image classification [6, 11], many existing approaches adopt a metric-based meta-learning paradigm, where videos are embedded into a shared feature space, and alignment metrics are leveraged to infer the labels of the queried video [3]. The core challenge lies in learning effective representations that can transfer to novel, unseen action categories, while overcoming the inherent complexity of video data and the limited supervision available [28, 39, 40, 54].

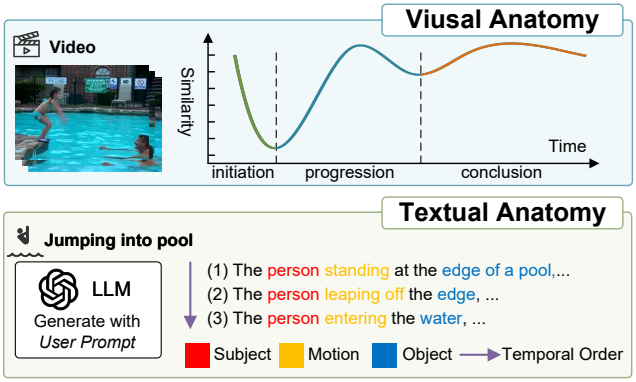


Figure 1. Illustration of our motivation. By leveraging the powerful knowledge understanding capability of LLMs, we anatomize one action label into a three-stage atomic action description. Meanwhile, the video is divided into corresponding three phases.

Recently, researchers have explored the integration of multimodal information to enhance the learned representations [7, 23, 34, 42, 43, 48]. Among these, text has emerged as a particularly effective modality [26, 34, 42]. For example, CLIP-FSAR [42] regards action labels as text and incorporates CLIP [30] to boost the FSAR performance. Shi *et al.* [34] constructs an external action knowledge base and aligns video frames with textual sub-action proposals. However, human actions exhibit intricate variations and rich spatiotemporal details [15, 21, 37, 55]. Subtle differences

*Corresponding author.

in posture, motion dynamics, and object interactions across different action phases serve as crucial knowledge for action recognition, yet these cues cannot be fully exploited by relying on simple semantic cues such as action labels.

To address this limitation, we first take a deeper look beyond the label semantics. Beyond the textual label, key factors of actions include the subject, motion dynamics, and object interactions [15, 32].Meanwhile, temporally, the way an action initiates, progresses, and concludes plays a crucial role in determining its category [8, 22, 53]. Thus, achieving robust few-shot action recognition requires fine-grained alignment between query and support videos, ensuring that every critical aspect of the action is considered.

With this insight, in this work, we propose **L**anguage-**G**uided Action **A**natomy (LGA): a novel framework that anatomizes both textual and visual modalities to fully exploit their inherent knowledge for fine-grained query-support matching. LGA first uses a Large Language Model (LLM) to decompose action labels into a sequence of semantic sub-steps, as illustrated in Fig. 1. We choose LLM due to its strong general world knowledge and instruction-following ability [25, 31], enabling us to focus on the core elements of actions, *i.e.*, subject, motion, and object. Simultaneously, a Visual Anatomy Module segments the video into distinct temporal phases, capturing the initiation, progression, and conclusion of the actions. A Fine-grained Multimodal Fusion Module then integrates the textual and visual features at the anatomized level, constructing modality-aligned action prototypes. Finally, to ensure robust few-shot matching, we propose a Multimodal Matching Module. This module performs matching from two complementary perspectives: video-video matching compares query and support video features, while video-text matching aligns query video features with textual representations. By fully decomposing and aligning the anatomized knowledge of actions, LGA enables fine-grained multimodal understanding and significantly improves few-shot action recognition performance.

Experimental results on multiple benchmarks demonstrate that our method outperforms other FSAR methods. Also, we conduct extensive ablation studies and provide qualitative visualizations, both of which substantiate the effectiveness of our method. The contributions can be summarized as follows:

- We propose LGA, a novel method for FSAR that anatomizes textual and video modalities to fully exploit the inherent knowledge in actions.
- We propose novel multimodal fusion and matching modules that effectively integrate atomic-level textual and visual features, enabling robust FSAR through both video-video and video-text matching.
- Our method achieves state-of-the-art performance across five widely used benchmarks.

2. Related work

2.1. Few-shot Action Recognition

Few-shot learning aims to develop models that can generalize to novel categories with only a limited number of labeled samples. Few-shot action recognition deals with videos that contain complex spatio-temporal information, which requires not only a deep understanding of contextual semantics, but also a comprehensive modeling of spatio-temporal relationships. Mainstream FSAR methods follow a "metric-based meta-learning" paradigm, where a common embedding space is first learned through episodic training, and alignment metrics are designed to calculate the similarities between the query video and support videos. For example, OTAM [3] uses dynamic time warping to better align the support and query videos temporally. TARN [1] and MPRE [24] both introduce temporal attention mechanisms to help the network generalize better to unseen categories. CPMT [13, 14] constructs multiple groups of prototypes to align each query against sequences in the support set and then fuses their matching scores for classification. HCL [54] introduces a hierarchical matching model to support comprehensive similarity measures at global, temporal, and spatial levels via a zoom-in matching module.

While these methods have explored various temporal modeling and alignment strategies, they are fundamentally constrained by their reliance on single-modality representations. This leads to overfitting on seen categories and limits generalization to novel action classes in FSAR. In our approach, we introduce the fine-grained fusion of visual and semantic features and multimodal classification between support videos and query videos, thereby significantly enhancing the model's performance in few-shot scenarios.

2.2. Multimdoal FSAR

To alleviate the data scarcity issue, several methods attempt to enhance feature representation with the complementarity of different modalities in FSAR. For example, AMeFu-Net [7] proposes to use depth information for FSAR. MoLo [43] and AMFAR[46] exploit the complementarity between visual features and motion cues. Furthermore, Lite-MKD [23] utilizes a teacher model to conduct multimodal learning to fuse optical flow, depth, and appearance features of human movements.

The integration of textual information shows great promise due to the high-quality representations provided by the pre-trained backbones such as CLIP [30]. CLIP-FSAR [42] and MORN [26] apply the pre-trained CLIP model to extract text and visual features, leveraging action labels as textual cues to enhance prototype learning. Inspired by this, several methods [2, 10, 20, 47, 49, 51] adopt adapter-based approaches to balance computational

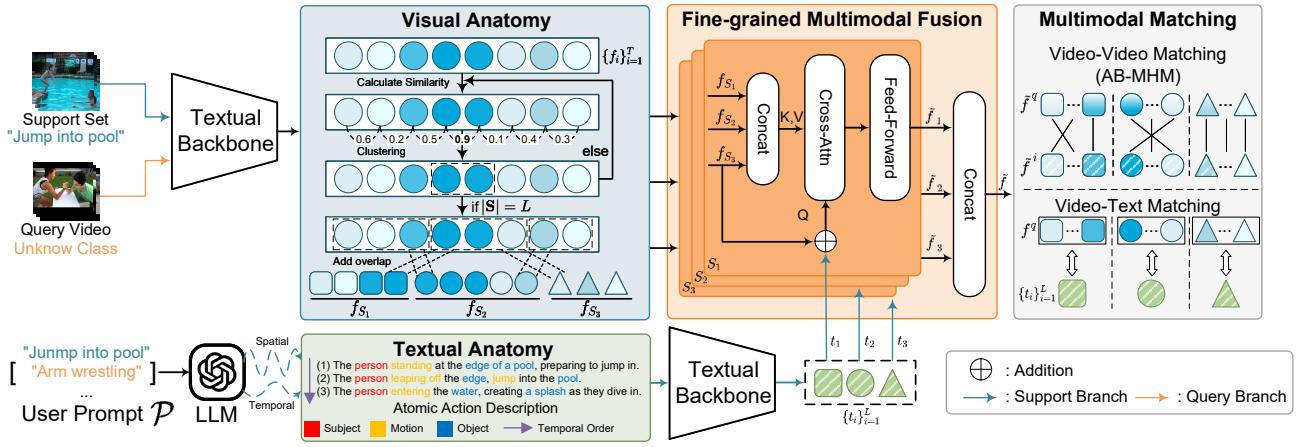


Figure 2. Illustration of our proposed method. First, the support and query videos are processed by the visual backbone to extract the visual features. Then, we transfer the action labels into atomic action descriptions with pre-trained large language model (LLM). Simultaneously, the Visual Anatomy Module segment frame sequence into atomic action phases. Then, we apply a Fine-grained Multimodal Fusion Module to integrate the visual features and semantic information at the atomic level, resulting in the final action prototype. Finally, we introduce a Multimodal Matching Module to achieve robust few-shot matching, which consists of both video-video matching and video-text matching.

efficiency and accuracy in FSAR. Furthermore, Xia *et al.* [48] and Shi *et al.* [34] construct external action knowledge bases from large-scale video captions and text corpora to enrich action representations. SAFSAR [38] extends elaborated descriptions from class names to better integrate textual and visual features in FSAR.

While these methods exploit multimodal cues, they fail to fully decompose and leverage the inherent structure of actions, leading to suboptimal prototype representations for novel action categories. In contrast, our proposed LGA framework addresses this gap with two key innovations: (1) Previous approaches rely solely on action labels or externally collected text proposals, failing to fully exploit the inherent knowledge in the text and video of the actions. In contrast, we decompose action labels into a sequence of atomic action steps that explicitly encode the three core elements of actions (subject, motion, and object). Correspondingly, we anatomize the video into temporal segments of initialization, progression, and conclusion, yielding finer-grained textual and visual representations. (2) Furthermore, our method aligns and fuses textual and visual features at the atomic level to generate robust prototypes, while previous CLIP-based FSAR methods neglect fine-grained alignment and fusion between different modalities.

3. Method

3.1. Overview

Problem Definition. The goal of few-shot action recognition (FSAR) is to classify an unlabeled query video into one of the N action classes in the "support set" comprising K labeled samples for each class that is unseen during training (N-way K-shot task). Following the typical "metric-based meta-learning " paradigm [35], we train a deep network on the training set $\mathcal{D}_{train}$ and then verify the model on the test set $\mathcal{D}_{test}$. All categories in the test set are not included in the training set. As in previous approaches [42, 43], in the training phase, we adopt the episode-based meta-learning strategy to optimize the network. In each episode, the network is trained to classify query samples into correct categories, and optimized to assume that query samples are in the same category as the support samples closest to it. In the test phase, a large number of episodes are randomly selected from the test set, and the average accuracy is utilized to evaluate the few-shot performance of the trained model.

Overall Architecture. An overview of our framework is provided in Fig. 2, which mainly consists of three components: Action Anatomy, Fine-grained Multimodal Fusion, and Multimodal Matching. Following previous methods [27, 43, 54], we first sparsely random sample T frame from the videos, and extract the image features with the feature extractor. We then anatomize actions in both textual and video modalities. Specifically, for text, we leverage pre-trained LLM to transform each label into sequences of atomic action descriptions, focusing on the three core elements of action: subject, action, and object. For videos, we decompose the visual features of the whole action into three sets of atomic action sequences, representing the initiation, progression, and conclusion phases. Then, to enhance the generalization ability of the action prototype, we use a Fine-grained Multimodal Fusion Module to integrate the anatomized visual and textual features, obtaining the action prototypes. Finally, based on the action prototypes, we leverage a novel Multimodal Matching Module to achieve robust few-shot matching. The Multimodal Matching consists of video-video matching and video-text matching. We

introduce the details of each component below.

3.2. Action Anatomy

The first key step in LGA is action anatomy, which is conducted on both the action labels and the videos. The anatomized text and videos are then fused and matched by subsequent modules for few-shot action recognition.

Textual Anatomy. To harness the rich prior knowledge embedded in LLM, we decompose each action label into an ordered sequence of atomic action descriptions that capture fine-grained spatiotemporal context. Specifically, we design a carefully crafted prompt that guides the LLM to decompose an action label into a sequence of semantically meaningful sub-actions, ensuring that the subject, motion, and object are explicitly represented at each stage. The generated descriptions follow the natural flow of action: from *initiation* to *progression* and finally to *conclusion*, allowing for a more structured and expressive textual representation.

An example of the anatomy of the label "Jump into pool" can be seen in Fig. 2. The LLM generates "A person standing at the edge of a pool, preparing to jump in.", "The person leaping off the edge, jump into the pool." and "The person entering the water, creating a splash as he dives in." These fine-grained atomic-level action descriptions provide explicit spatiotemporal cues compared to the action label. Finally, we feed the atomic action descriptions into the textual backbone to extract features $\{t_i\}_{i=1}^{L}$, where L is the number of the generated atomic descriptions.

Visual Anatomy: The Visual Anatomy Module transforms sequences of frame features into sets of atomic phases corresponding to the generated atomic action descriptions. As shown in Fig. 2, given a video sequence $\{f_i\}_{i=1}^{T}$, we transform it into distinct groups of atomic action sequences $\{f_{S_1}, f_{S_2}, ..., f_{S_L}\}$, where L represents the number of atomic actions in the video, corresponding to the atomic action textual descriptions.

To accurately anatomize the visual sequence feature into atomic phases, we propose a CLUSTER-Segment strategy that identifies sub-actions of varying temporal durations. This strategy encourages the frames within the same sub-action to be similar, while ensuring that frames belonging to different sub-actions remain as distinct as possible. Specifically, for an initial frame sequence $\{f_i\}_{i=1}^{T}$, we treat each frame as an individual cluster $\{S_i\}_{i=1}^{T}$. We then compute the cosine similarity between adjacent clusters, and merge the pair of adjacent clusters exhibiting the highest similarity to obtain the updated cluster $\{S_i\}_{i=1}^{T-1}$. This merging process is repeated until only L clusters remain. Additionally, to enhance the robustness of the CLUSTER-Segment algorithm, we introduce duplicated frames within adjacent clusters. More details are described in Algorithm 1.

Algorithm 1 Algorithm for CLUSTER-Segment in Visual Anatomy Module

Require: frame feature $\{f_i\}_{i=1}^{T}, f_i \in \mathbb{R}^C$
  initialize cluster $\mathbf{S} = \{S_i\}_{i=1}^{T}$,
  target number of atomic actions of L
  while $|\mathbf{S}| > L$ **do**
    $\mathbf{L}_{sim}$ = []
    $\overline{f}_S = \frac{1}{|S_i|} \sum f_i \quad$ where $i \in S_i$
    for $i = 0$ **to** $|\mathbf{S}|$ **do**
      Sim = CosineSimilarity$(\overline{f}_{S_i}, \overline{f}_{S_{i+1}})$
      # Append Sim to $\mathbf{L}_{sim}$
    end for
    j = argmax$(\mathbf{L}_{sim})$
    # Merge cluster $\mathbf{S_j}$ and $\mathbf{S_{j+1}}$
  end while
  # Add overlap frame within adjacent clusters
  return Temporal-Segment Result $\mathbf{S} = \{S_i\}_{i=1}^{L}$

3.3. Fine-grained Multimodal Fusion

After anatomizing both visual and textual features, the next challenge is to effectively integrate these two modalities to construct a robust action representation. To this end, we propose a Fine-grained Multimodal Fusion Module designed to align and fuse visual and semantic information at the anatomized action level, as illustrated in Fig. 2.

In Fine-grained Multimodal Fusion Module, given the anatomized visual feature $\{f_{S_i} \mid S_i \in \{S_1, \ldots, S_L\}\}$ and its corresponding textual feature $\{t_i\}_{i=1}^{L}$, our goal is to establish fine-grained correspondences between these representations. Specifically, we employ a multi-head Cross-Attention mechanism to facilitate interaction between modalities. In our design, we formulate the query $\mathbf{Q}$ as the sum of the atomic visual feature f_{S_i} and its aligned textual feature t_i. Meanwhile, the key and value matrices $\mathbf{K}$ and $\mathbf{V}$ are constructed by concatenating all atomic visual features. This design ensures that each atomic action feature not only learns localized semantic details but also maintains awareness of the global temporal structure of the action. Subsequently, a feed-forward network processes the cross-attended information to generate the fused feature $\tilde{f}_{S_i}$ for the atomic action. Finally, we concatenate the atomic action features across all temporal phases to construct the final action prototype $\tilde{f}$. The process can be formulated as:

$$\begin{aligned} &\mathbf{Q_i} = t_i + f_{S_i}, \mathbf{K} = \mathbf{V} = \text{concat}(\{f_{S_i}\}_{i=1}^{L}), \\ &f_{S_i} = \text{Softmax}\left(\frac{\mathbf{Q_i}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tilde{f}_{S_i} = \mathbf{FFN}(f_{S_i}) + f_{S_i}, \\ &\tilde{f} = \text{concat}(\tilde{f}_{S_1}, \tilde{f}_{S_2}, \tilde{f}_{S_3}). \end{aligned} \quad (1)$$

We then show how we leverage the action prototype for robust few-shot matching.

3.4. Multimodal Matching

In few-shot scenarios, accurate classification depends on more than just measuring the distance between query and support videos based solely on visual feature matching. Previous methods often overlook the use of multimodal information, which limits their ability to handle the richness of both visual and textual cues [27, 52]. Leveraging our anatomized representations and fused prototypes, we propose a Multimodal Matching Module that enhances classification accuracy by considering both video-video matching and video-text matching.

Video-video matching: In video-video matching, we measure the distance between the features of the query video and the prototype features of each category within the support set. In this context, it is crucial to account for the ordered temporal alignment between atomic actions. Existing methods, such as the bidirectional Mean Hausdorff Metric (Bi-MHM) [40], typically operate on a global temporal scale and do not incorporate the inherent order of atomic actions across the video sequence. This approach fails to recognize that actions consist of temporally ordered sub-actions, and the distance between identical actions should be evaluated based on the alignment of these sub-actions at each time step.

To address this gap, we propose a novel metric **A**ligned **B**idirectional **M**ean **H**ausdorff **M**etric (AB-MHM), which explicitly incorporates ordered temporal alignment of atomic actions:

$$\begin{aligned}\tilde{\mathcal{D}}(\tilde{f}^q, \tilde{f}^i) = \frac{1}{T}\sum_{k=1}^{L}[\sum_{\tilde{f}_m^q \in \tilde{f}_{S_k}^q} (\min_{\tilde{f}_n \in \tilde{f}_{S_k}} \left|\left| \tilde{f}_m^q - \tilde{f}_n^i \right|\right|) \\ + \sum_{\tilde{f}_n \in \tilde{f}_{S_k}} (\min_{\tilde{f}_m^q \in \tilde{f}_{S_k}^q} \left|\left| \tilde{f}_n^i - \tilde{f}_m^q \right|\right|)],\end{aligned} \tag{2}$$

$$p_{(y=i|q)}^{v-v} = \text{Softmax}(-\tilde{\mathcal{D}}(\tilde{f}^q, \tilde{f}^i)), \tag{3}$$

where T denotes the frame number. Finally, we take the negative distance for each class as the logit, and employ a softmax function to derive the probability p^{v-v} that the query video belongs to category, thereby obtaining the classification results for video-video matching.

The key distinction of AB-MHM is that it aligns atomic actions in time, ensuring a more accurate measure of similarity. This approach not only improves the accuracy of video-video matching but also increases computational efficiency. AB-MHM is non-parametric and does not require complex alignment classifiers, making it more transferable and less computational costly. The benefits of incorporating ordered temporal alignment are demonstrated in Sec. 4.3, where our method shows significant improvements over traditional matching approaches.

Video-text Matching: In video-text matching, we classify the query video by measuring the distance between its feature representation and the atomic-level textual features of each category. Specifically, given a query feature $\tilde{f}^q$, we first apply average pooling to obtain the feature in each action phase $\overline{f}_{S_k}^q$. Then, we calculate the similarity between the visual and semantic features at each action phase and classify the query video using a softmax function, which can be expressed as:

$$p_{(y=i|q)}^{v-t} = \text{Softmax}(\sum_{k=1}^{L} \langle \overline{f}_{S_k}^q, f_k^T \rangle)) \tag{4}$$

Finally, we integrate the classification results from video-video matching and video-text matching using a weighted geometric mean as in [42], which can be formulated as:

$$p_{(y=i|q)} = \left(p_{(y=i|q)}^{v-v}\right)^{\alpha} \times \left(p_{(y=i|q)}^{v-t}\right)^{(1-\alpha)} \tag{5}$$

where $\alpha \in [0, 1]$ is an adjustable hyperparameter.

During the training phase, the entire framework is trained end-to-end, and we utilize the cross-entropy loss [3, 27] and contrastive loss [43, 45] to optimize the model. The contrastive loss enforces consistency in feature representations by contrasting positive and negative samples. Notably, to ensure training stability, we adopt the results from video-video matching as the classification results during the training phase. During inference, we adopt the results of Multimodal Matching as the classification results for the query video.

4. Experiments

4.1. Experimental setup

Datasets: We perform our experiments on five widely used few-shot action recognition datasets: SSv2-Full [9], SSv2-Small [9], Kinetics [4], UCF101 [36], and HMDB51 [17]. For a fair comparison with previous methods, we follow the dataset split as described in [3, 56].

Training and Inference: Following previous methods [27, 40, 42, 43], we uniformly sample 8 frames (*i.e.*, $T = 8$) to represent each video, and employ a ViT-B/16 [5] backbone with CLIP [30] initialization to serve as the visual and textual encoder. In the training stage, our method is optimized end-to-end using the Adam optimizer [16]. In the inference stage, we centrally crop a 224×224 region from each frame. Following previous methods [39, 40, 43], we calculate the average classification accuracy of the 10,000 episodes from the test set to evaluate the few-shot performance on each benchmark.

4.2. Comparison with state-of-the-art methods

In this section, we use CLIP-FSAR [42] as a baseline method and comprehensively compare our method with

Table 1. Comparison with state-of-the-art few-shot action recognition methods on HMDB51, Kinetics, UCF101, SSv2-Full, and SSv2-Small datasets with 5-way k-shot setting. The best results are **bolded** in black and the underline represents the second best result.

Method	Reference	Pre-training	HMDB51		Kinetics		UCF101		SSv2-Small		SSv2-Full	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
OTAM [3]	CVPR'20	INet-RN50	54.5	68.0	73.0	85.8	79.9	88.9	36.4	48.0	42.8	52.3
TRX [27]	CVPR'21	INet-RN50	53.1	75.6	63.6	85.9	78.2	96.1	36.0	56.7	42.0	64.6
STRM [39]	CVPR'22	INet-RN50	52.3	77.3	62.9	86.7	80.5	96.9	37.1	55.3	43.1	68.1
HyRSM [40]	CVPR'22	INet-RN50	60.3	76.0	73.7	86.1	83.9	94.7	40.6	56.1	54.3	69.0
CMPT [13]	ECCV'22	INet-RN50	60.1	77.0	73.3	86.4	71.4	91.0	38.9	61.6	49.3	66.7
MoLo [43]	CVPR'23	INet-RN50	60.8	77.4	74.0	85.6	86.0	95.5	42.7	56.4	56.6	70.6
GgHM [50]	ICCV'23	INet-RN50	61.2	76.9	74.9	87.4	85.2	96.3	-	-	54.5	69.2
Huang *et al.* [14]	IJCV'24	INet-RN50	61.6	77.5	74.0	86.9	74.9	92.5	42.6	61.8	52.3	67.1
TEAM [19]	CVPR'25	INet-RN50	62.8	78.4	75.1	88.2	87.2	96.2	-	-	-	-
OTAM [3]	CVPR'20	CLIP-ViT-B	72.5	83.9	88.2	94.8	95.8	98.8	43.3	57.5	50.2	68.6
CLIP-Freeze [30]	ICML'21	CLIP-ViT-B	58.2	77.0	78.9	91.9	89.7	95.7	29.5	42.5	30.0	42.4
CapFSAR [44]	ArXiv'23	BLIP-ViT-B	65.2	78.6	84.9	93.1	93.3	97.8	45.9	59.9	51.9	68.2
MVP-shot [29]	Arxiv'24	CLIP-ViT-B	77.0	88.1	91.0	95.1	96.8	99.0	55.4	62.0	-	-
EMP-Net [47]	ECCV'24	CLIP-ViT-B	76.8	85.8	89.1	93.5	94.3	98.2	57.1	65.7	63.1	73.0
CLIP-CPM^{2}C [10]	Nc'25	CLIP-ViT-B	75.9	88.0	91.0	95.5	95.0	98.6	52.3	62.6	60.1	72.8
CLIP-FSAR [42]	IJCV'23	CLIP-ViT-B	77.1	87.7	94.8	95.4	97.0	**99.1**	54.6	61.8	62.1	72.1
Our	-	CLIP-ViT-B	**86.8**	**89.3**	**95.2**	**96.2**	**98.2**	**99.1**	**58.9**	**69.3**	**63.8**	**74.4**

state-of-the-art approaches [13, 14, 40, 42, 43, 47, 54] using mainstream backbones (INet-RN50 [12] or ViT-B/16 [5]). For a fair comparison, we exclude task-specific methods [2, 41] that leak task information, which compromise the transferability of the model.

As shown in Tab. 1, we can make the following observations: (a) Compared to the INet-RN50 [12] backbone, the methods with the CLIP-ViT-B [5] backbone show superior performance. This indicates that the transformer architecture and pre-training contribute significantly to the performance. (b) Compared to the baseline (CLIP-FSAR [42]), our method achieves superior performance across all datasets, demonstrating its effectiveness in different scenarios. (c) Compared to other methods, our approach also achieves competitive performance. Notably, the experiments show that our method delivers larger performance improvements on Kinetics and HMDB51. We attribute this to the fact that, compared to the SSv2 dataset, these datasets contain rich action-related elements that align well with our atomic action descriptions. For UCF101, the training data of the backbone already cover the action categories, leaving little room for further improvement.

4.3. Ablation study

Analysis of network components. In Tab. 2, we perform ablation studies in 5-way 1/5-shot settings to investigate the effectiveness of different components in our methods. We can find that anatomizing the action and its label into atomic units significantly improves the performance of FSAR, improving 1-shot accuracy on HMDB51 by 4.1% (79.9% vs. 75.8%) and 3.8% (79.6% vs. 75.8%), respectively. In addition, the application of action decomposition in both the visual and textual domains improves performance by a further 5.0%. This consistent improvements underscore the importance and effectiveness of atomic-level action decomposition for FSAR. However, we observe that using visual or textual anatomy alone leads to performance degradation in the 5-shot setting. We attribute this to the fact that in the 5-shot setting, sufficient samples in the support set allow the model to achieve effective classification. However, the addition of unimodal anatomy may introduce misalignment between different modalities, causing overfitting and degraded performance. Furthermore, the additional video-text matching in our multimodal matching yields improvements of 6.0% and 1.1% under 1-shot and 5-shot settings, respectively, on HMDB51, confirming the effectiveness of multimodal matching. Notably, the performance increase is particularly significant in the 1-shot scenario, highlighting

Table 2. Ablation study on HMDB51 datasets under 5-way 1-shot and 5-way 5-shot settings. The top two line represents the baseline CLIP-FSAR. "Visual-An" and "Textual-An" mean anatomize the video and action label into atomic units, respectively.

Fusion		Macthing		HMDB51	
Visual-An	Textual-An	video-video	video-text	1-shot	5-shot
		✓		75.8	87.7
		✓	✓	77.1	87.7
✓		✓		79.9	86.0
	✓	✓		79.6	87.2
✓	✓	✓		80.8	88.2
✓		✓	✓	83.1	86.2
	✓	✓	✓	85.2	87.8
✓	✓	✓	✓	**86.8**	**89.3**

Table 3. Comparison of different multimodal fusion schemes on the HMDB51 and SSv2-Small datasets with 5-way 1/5-shot settings. "G" and "L" denote global and local multimodal fusion between visual features and semantic cues, respectively.

Text	Scale	Attention	HMDB51		SSv2-Small	
			1-shot	5-shot	1-shot	5-shot
Action Label	G	Self	79.3	85.9	58.0	66.1
Action Label	G	Cross	81.7	85.9	58.5	67.2
Atomic Description	G	Self	85.2	87.6	58.4	65.4
Atomic Description	L	Self	85.5	88.8	57.8	66.5
Atomic Description	L	Cross	86.8	89.3	58.9	69.3

Table 4. Comparison results of different few-shot matching strategies on the HMBD51 and SSv2-Small datasets.

Matching		HMDB51		SSv2-Small	
video-video	video-text	1-shot	5-shot	1-shot	5-shot
OTAM [3]	-	80.9	88.1	55.2	66.3
Bi-MHM [40]	-	81.1	88.2	54.8	66.1
AB-MHM	-	80.9	88.2	56.5	69.2
-	Action label	72.5		35.6	
-	Atomic description	81.9		36.1	
OTAM [3]	Atomic description	86.7	89.3	56.5	66.3
Bi-MHM [40]	Atomic description	87.6	89.6	55.2	66.4
AB-MHM	Atomic description	86.8	89.3	58.9	69.3

the effectiveness of decomposing actions into atomic units and few-shot matching with different modality cues when visual information is limited. Moreover, the best performance is achieved by using both components together, empirically validating that multimodal fusion and multimodal matching complement each other.

Analysis of Multi-modal Fusion. To better demonstrate the effectiveness of our Fine-grained Multimodal Fusion Module, we explore different multimodal fusion schemes to integrate visual features and semantic information. As shown in the Tab. 3, atomic action descriptions outperform action labels within the same fusion framework, as they provide more effective semantic information about the action. In addition, the cross-attention mechanism efficiently exploits semantic information by computing the similarity between textual and visual modality features, allowing it to better align and fuse the prior knowledge across different modalities. This enables the construction of more effective action representations in few-shot scenarios. Overall, compared to the global-scale fusion of visual and textual features in CLIP-FSAR [42], our method integrates textual and visual features at a fine-grained level, thus achieving more efficient action prototype construction and better few-shot action recognition results.

Analysis of matching mechanism. In this section, we explore the role of our Multimodal Matching Module within the LGA framework. As shown in Tab. 4, we perform ablation experiments by changing matching strategies on the HMDB51 and SSv2-Small datasets. On HMDB51, the performance of different video-video matching strategies is generally comparable, but the additional atomic descriptions in our method significantly improve the video-text matching branch compared to the action label. Conversely, on SSv2-Small, video-text matching yields limited performance gains, while our proposed AB-MHM outperforms the Bi-MHM method branch by 3.7% and 2.9% under the 1-shot and 5-shot settings, respectively. We attribute this to the fact that HMDB51 contains richer scene information, enabling more accurate classification through enriched textual cues, while SSv2-Small has a stronger temporal focus with less scene content, making atomic-based video-video matching more effective. Overall, these results demonstrate that our Multimodal Matching Module delivers robust few-shot matching across different datasets and settings.

Table 5. Comparison of different Temporal Segment methods on the HMDB51 and SSv2-Small datasets.

Segment Method	HMDB51		SSv2-Small	
	1-shot	5-shot	1-shot	5-shot
TW-FINCH [33]	82.9	86.2	56.4	66.2
HARD	**87.3**	**89.3**	57.0	**69.3**
CLUSTER w/o overlap	86.6	88.3	58.3	68.0
CLUSTER	86.8	**89.3**	**58.9**	**69.3**

Analysis of segment methods in Visual Anatomy. In this section, we analyze the effect of different temporal segmentation strategies in our framework. TW-FINCH [33] is a temporally-weighted hierarchical clustering algorithm based on a 1-nearest neighbor graph. We use another baseline HARD that uniformly divides the video into three segments of equal length with the same overlap as ours. As shown in Tab. 5, we observe that our CLUSTER-Segment method consistently outperforms other competitors. Specifically, TW-FINCH faces challenges in FSAR due to the lack of clear feature boundaries between atomic actions. Furthermore, compared to HARD and CLUSTER (without overlap), our proposed CLUSTER—Segment method better captures atomic actions of varying temporal lengths. This enables the construction of reliable prototypes based on atomic actions, leading to improved performance.

Analysis of Atomic Action number. We analyze the impact of varying the number of atomic actions on model performance. As shown in Tab. 6, performance improves as the number of atomic actions increases, peaking at 3, but declines when increased to 4. This suggests a balance between granularity and coherence: while finer segmentation enhances action understanding, excessive splitting may introduces excessive temporal overlap, which dilutes the core motion signals and degrades model performance. Additionally, a larger number of atomic actions amplifies the risk of LLM hallucinations, leading to ambiguous descriptions and misalignment between visual and textual features (further

Table 6. Analysis of different the num of atmoic action on the HMDB51 and SSv2-Small datasets.

Number	HMDB51		SSv2-Small	
	1-shot	5-shot	1-shot	5-shot
1	54.3	65.4	86.7	88.7
2	57.5	67.3	87.6	89.3
3	58.9	69.3	86.8	89.3
4	53.6	62.7	73.9	82.5

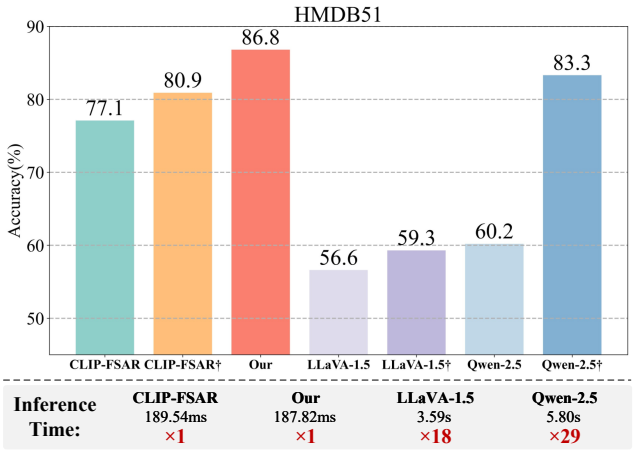


Figure 3. Comparison experiments on the effect of LLM and VLM on the HMDB51 dataset.

analysis in Sec. 8.2). Therefore, we default the number of atomic actions to 3, which aligns well with the initiation, progression, and conclusion phases of an action.

Analysis of pre-trained LLM and VLM In this section, we explore the impact of atomic action descriptions generated by LLM on existing FSAR methods. Specifically, we replace the action labels in CLIP-FSAR [42] with atomic action descriptions generated by the LLM model, resulting in CLIP-FSAR†. As shown in Fig. 3, the use of atomic action descriptions improves the performance of CLIP-FSAR, indicating that enriched action descriptions effectively improve the FSAR model's ability to represent actions in few-shot scenarios. In addition, we investigate the performance of pre-trained VLMs in the few-shot action recognition task. Specifically, we evaluate the LLaVA-1.5-7B and Qwen-2.5-7B models (details in supplementary). As shown in Fig. 3, Qwen-2.5 slightly outperforms LLaVA-1.5 in zero-shot settings, but its performance remains significantly below that of mainstream FSAR methods. Furthermore, we explore the incorporation of label information into the prompt to obtain LLaVA-1.5† and Qwen-2.5†, enabling further evaluation of VLM performance. Although Qwen-2.5† outperforms most mainstream FSAR methods, the high computational requirements of VLM networks lead to significantly increase inference time and hinder its scalability across different N-way K-shot settings. In contrast, the proposed LGA effectively leverages LLMs to explore action prior knowledge, thereby improving model performance while maintaining high inference efficiency.

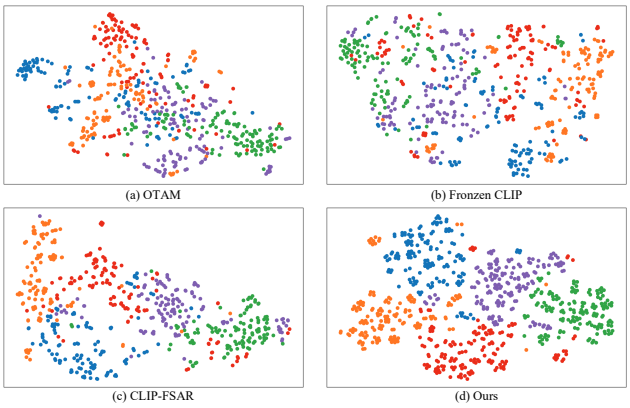


Figure 4. T-SNE distribution visualization of five action classes on the test set of SSv2-Small dataset. The different color represents video from different categories.

4.4. Visualizations

Visualization of Prototype Similarity. To further analyze our method, we visualize the feature distribution of the baseline methods [3, 30, 42] and our method in the test phase. As shown in Fig. 4, we observe that our approach significantly improves the discriminative capabilities of action video feature representations compared to previous methods. Specifically, due to the limited training data and temporal dynamic information in the SSv2-Small dataset, OTAM [3] and Frozen-CLIP [30] fail to produce accurate feature distributions for different action categories. While CLIP-FSAR [42] shows some improvement, demonstrating the effectiveness of incorporating multimodal information in few-shot action recognition, it still suffers from significant feature distribution overlap. In contrast, our method effectively captures discriminative features for different actions, resulting in a more accurate feature distribution and improved prediction accuracy.

5. Conclusion

We introduce Language-Guided Action Anatomy (LGA), a novel framework for few-shot action recognition that models actions at a fine-grained level by anatomizing both visual and textual modalities. Unlike prior methods that rely on coarse action labels, LGA structures actions into atomic phases, capturing subject, motion, and object interactions during the initiation, progression, and conclusion of the action. We integrate textual and visual features at the atomic level to generate more generalizable prototypes, and we present a Multimodal Matching Module for robust few-shot matching. Extensive experiments on five widely used benchmarks demonstrate the effectiveness of our approach, with LGA achieving state-of-the-art performance.

6. Acknowledge

This work was partly funded by the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and STCSM (22DZ2229005).

References

[1] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. TARN: temporal attentive relation network for few-shot and zero-shot action recognition. In *BMVC*, page 154, 2019. 2

[2] Congqi Cao, Yueran Zhang, Yating Yu, Qinyi Lv, Lingtong Min, and Yanning Zhang. Task-adapter: Task-specific adaptation of image models for few-shot action recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9038–9047, 2024. 2, 6, 4

[3] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *CVPR*, pages 10618–10627, 2020. 1, 2, 5, 6, 7, 8, 3, 4

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 5

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 5, 6

[6] Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 594–611, 2006. 1

[7] 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. 1, 2

[8] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Fine-grained temporal contrastive learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19999–20009, 2022. 2

[9] 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. 5

[10] Fei Guo, YiKang Wang, Han Qi, Li Zhu, and Jing Sun. Consistency prototype module and motion compensation for few-shot action recognition (clip-cpm2c). *Neurocomputing*, 611:128649, 2025. 2, 6, 4

[11] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE international conference on computer vision*, pages 3018–3027, 2017. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[13] Yifei Huang, Lijin Yang, and Yoichi Sato. Compound prototype matching for few-shot action recognition. In *European Conference on Computer Vision*, pages 351–368. Springer, 2022. 2, 6

[14] Yifei Huang, Lijin Yang, Guo Chen, Hongjie Zhang, Feng Lu, and Yoichi Sato. Matching compound prototypes for few-shot action recognition. *International Journal of Computer Vision*, pages 1–26, 2024. 2, 6, 3

[15] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020. 1, 2

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[17] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 5, 3, 4

[18] Pulkit Kumar, Namitha Padmanabhan, Luke Luo, Sai Saketh Rambhatla, and Abhinav Shrivastava. Trajectory-aligned space-time tokens for few-shot action recognition. In *European Conference on Computer Vision*, pages 474–493. Springer, 2024. 4

[19] SuBeen Lee, WonJun Moon, Hyun Seok Seong, and Jae-Pil Heo. Temporal alignment-free video matching for few-shot action recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5412–5421, 2025. 6

[20] Bozheng Li, Mushui Liu, Gaoang Wang, and Yunlong Yu. Frame order matters: A temporal sequence-aware model for few-shot action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18218–18226, 2025. 2, 4

[21] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 382–391, 2020. 1

[22] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3889–3898, 2019. 2

[23] Baolong Liu, Tianyi Zheng, Peng Zheng, Daizong Liu, Xiaoye Qu, Junyu Gao, Jianfeng Dong, and Xun Wang. Litemkd: A multi-modal knowledge distillation framework for lightweight few-shot action recognition. In *Proceedings*

of the 31st ACM International Conference on Multimedia, pages 7283–7294, 2023. 1, 2

[24] Shuwen Liu, Min Jiang, and Jun Kong. Multidimensional prototype refactor enhanced network for few-shot action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6955–6966, 2022. 2

[25] Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1:3, 2020. 2

[26] Xinzhe Ni, Yong Liu, Hao Wen, Yatai Ji, Jing Xiao, and Yujiu Yang. Multimodal prototype-enhanced network for few-shot action recognition. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1–10, 2024. 1, 2

[27] 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. 3, 5, 6

[28] Zefeng Qian, Chongyang Zhang, Yifei Huang, Gang Wang, and Jiangyong Ying. Joint image-instance spatial-temporal attention for few-shot action recognition. *Computer Vision and Image Understanding*, page 104322, 2025. 1

[29] Hongyu Qu, Rui Yan, Xiangbo Shu, Hailiang Gao, Peng Huang, and Guo-Sen Xie. Mvp-shot: Multi-velocity progressive-alignment framework for few-shot action recognition. *arXiv preprint arXiv:2405.02077*, 2024. 6

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 5, 6, 8, 4

[31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 2

[32] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11184–11193, 2021. 2

[33] Saquib Sarfraz, Naila Murray, Vivek Sharma, Ali Diba, Luc Van Gool, and Rainer Stiefelhagen. Temporally-weighted hierarchical clustering for unsupervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11225–11234, 2021. 7

[34] Yuheng Shi, Xinxiao Wu, Hanxi Lin, and Jiebo Luo. Commonsense knowledge prompting for few-shot action recognition in videos. *IEEE Transactions on Multimedia*, 2024. 1, 3

[35] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *NeurIPS*, 30, 2017. 3

[36] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[37] Li Sun, Yanjun Chen, Wu Luo, Haiyan Wu, and Chongyang Zhang. Discriminative clip mining for video anomaly detection. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2121–2125. IEEE, 2020. 1

[38] Yutao Tang, Benjamín Béjar, and René Vidal. Semantic-aware video representation for few-shot action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6458–6468, 2024. 3

[39] 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. 1, 5, 6

[40] 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. *2022CVPR*, 2022. 1, 5, 6, 7, 3

[41] Xiao Wang, Weirong Ye, Zhongang Qi, Guangge Wang, Jianping Wu, Ying Shan, Xiaohu Qie, and Hanzi Wang. Task-aware dual-representation network for few-shot action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10):5932–5946, 2023. 6

[42] Xiang Wang, Shiwei Zhang, Jun Cen, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong Sang. Clip-guided prototype modulating for few-shot action recognition. *International Journal of Computer Vision (IJCV)*, 2023. 1, 2, 3, 5, 6, 7, 8, 4

[43] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong Sang. 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*, pages 18011–18021, 2023. 1, 2, 3, 5, 6

[44] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Yingya Zhang, Changxin Gao, Deli Zhao, and Nong Sang. Few-shot action recognition with captioning foundation models. *arXiv preprint arXiv:2310.10125*, 2023. 6, 4

[45] Xiao Wang, Yan Yan, Hai-Miao Hu, Bo Li, and Hanzi Wang. Cross-modal contrastive learning network for few-shot action recognition. *IEEE Transactions on Image Processing*, 2024. 5

[46] 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. 2

[47] Cong Wu, Xiao-Jun Wu, Linze Li, Tianyang Xu, Zhenhua Feng, and Josef Kittler. Efficient few-shot action recognition via multi-level post-reasoning. In *European Conference on Computer Vision*, pages 38–56. Springer, 2025. 2, 6, 4

[48] Haifeng Xia, Kai Li, Martin Renqiang Min, and Zhengming Ding. Few-shot video classification via representation fusion

and promotion learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19311–19320, 2023. 1, 3

[49] Jiazheng Xing, Mengmeng Wang, Xiaojun Hou, Guang Dai, Jingdong Wang, and Yong Liu. Multimodal adaptation of clip for few-shot action recognition. *arXiv preprint arXiv:2308.01532*, 2023. 2

[50] Jiazheng Xing, Mengmeng Wang, Yudi Ruan, Bofan Chen, Yaowei Guo, Boyu Mu, Guang Dai, Jingdong Wang, and Yong Liu. Boosting few-shot action recognition with graph-guided hybrid matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1740–1750, 2023. 6

[51] Jiazheng Xing, Jian Zhao, Chao Xu, Mengmeng Wang, Guang Dai, Yong Liu, Jingdong Wang, and Xuelong Li. Ma-fsar: Multimodal adaptation of clip for few-shot action recognition. *Pattern Recognition*, page 111902, 2025. 2, 4

[52] Songyang Zhang, Jiale Zhou, and Xuming He. Learning implicit temporal alignment for few-shot video classification. *arXiv preprint arXiv:2105.04823*, 2021. 5

[53] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2914–2923, 2017. 2

[54] Sipeng Zheng, Shizhe Chen, and Qin Jin. Few-shot action recognition with hierarchical matching and contrastive learning. In *ECCV*. Springer, 2022. 1, 2, 3, 6

[55] Hao Zhou, Chongyang Zhang, Yan Luo, Chuanping Hu, and Wenjun Zhang. Thinking inside uncertainty: Interest moment perception for diverse temporal grounding. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):7190–7203, 2022. 1

[56] 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. 5