Trokens: Semantic-Aware Relational Trajectory Tokens for Few-Shot Action Recognition

Pulkit Kumar[1*] Shuaiyi Huang[1*]
Mathew Walmer[1] Sai Saketh Rambhatla[1,2] Abhinav Shrivastava[1]
[1]University of Maryland, College Park [2]GenAI, Meta
{pulkit,huangshy,mwalmer,abhinav}@cs.umd.edu rssaketh@meta.com

Abstract

Video understanding requires effective modeling of both motion and appearance information, particularly for few-shot action recognition. While recent advances in point tracking have been shown to improve few-shot action recognition, two fundamental challenges persist: selecting informative points to track and effectively modeling their motion patterns. We present Trokens, a novel approach that transforms trajectory points into semantic-aware relational tokens for action recognition. First, we introduce a semantic-aware sampling strategy to adaptively distribute tracking points based on object scale and semantic relevance. Second, we develop a motion modeling framework that captures both intra-trajectory dynamics through the Histogram of Oriented Displacements (HoD) and inter-trajectory relationships to model complex action patterns. Our approach effectively combines these trajectory tokens with semantic features to enhance appearance features with motion information, achieving state-of-the-art performance across six diverse few-shot action recognition benchmarks: Something-Something-V2 (both full and small splits), Kinetics, UCF101, HMDB51, and FineGym.

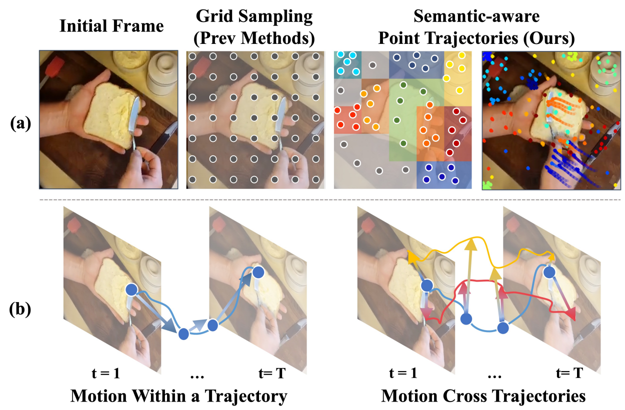

Figure 1. **Our motivation.** (a) Our semantic-aware points adapts better to object scale and semantic relevance while existing methods w/ grid sampling missed small objects w/ important motion (e.g., knife). (b) We explicitly model relational motions within a trajectory and across trajectories.

1. Introduction

At the core of video understanding lies the fundamental synergy between motion and appearance cues. Motion patterns reveal the dynamic flow of actions through time. At the same time, appearance information captures the rich context, the interplay of objects, environments, and their relationships within each frame. In action recognition tasks, particularly in few-shot settings, explicitly modeling this complementary relationship becomes crucial, as both aspects provide distinct yet essential signals.

While appearance understanding in models has made great strides, the challenge of capturing crucial motion patterns remains complex. Traditional optical flow techniques [30, 38, 44] have been a primary approach, but they are fundamentally limited to analyzing adjacent frames and deteriorate under occlusions, resulting in incomplete motion representations. In parallel to optical flow methods, trajectory-based approaches emerged as an alternative paradigm. Early trajectory works [50, 51, 53] made progress by capturing longer-term patterns, and recent work [28] has further advanced this direction through point tracking. Unlike optical flow, point tracking [10, 25] explicitly maintains temporal correspondence across long sequences and handles occlusions naturally, making it particularly effective for capturing complex motion patterns in real-world actions. In this work, we aim to advance few-shot action recognition by building upon these advantages of point tracking approaches.

To advance point tracking for action recognition, we must address two fundamental challenges: (1) sampling informative query points to track through time, and (2) effec-

[*]Equal Contribution

tively modeling the complex motion patterns captured by these trajectories. Dense point sampling provides comprehensive coverage but is computationally expensive, while sparse sampling risks missing crucial motion information, especially from smaller objects essential for action understanding. Beyond these tracking challenges, current transformer-based approaches [2, 3, 36] attempt to learn motion patterns implicitly for action recognition, but it remains unclear whether these models truly capture and leverage motion information or rely on other contextual cues.

To address these limitations, we identify two key questions: (1) how can we develop an effective sampling strategy that balances coverage and efficiency? and (2) how can we explicitly model and utilize the motion patterns captured in point trajectories? In this paper, we present a novel approach that leverages both semantic-aware sampling and explicit motion modeling to improve point tracking-based video understanding.

For the first challenge of effective point sampling, we propose a semantic-aware sampling strategy that adapts to object scale and importance. This approach is particularly crucial for actions involving small but critical objects, as illustrated in Figure 1(a), where the knife spreading the butter could be easily missed by uniform sampling due to its small size. By leveraging semantic information extracted from DINO [35] patch tokens, our method ensures comprehensive coverage of action-relevant objects regardless of scale. The sampling density, guided by semantic understanding, allocates denser points to smaller, action-critical objects (*e.g.,* knife) and sparser sampling to larger regions (*e.g.,* background desk). This approach ensures we capture the motion of all semantically meaningful objects while maintaining computational efficiency.

With our sampling strategy in place, we address the second challenge through our Relational Motion Modeling module, which processes point trajectories from modern trackers [10, 25] in two complementary ways. Revisiting the knife spreading butter example in Figure 1, our first component captures intra-trajectory dynamics using Histogram of Oriented Displacements (HoD) [17] to model the knife's own movement patterns and directions. Our second component extracts inter-trajectory relationships by tracking how different objects (*e.g.,* the knife and bread) interact with each other, revealing the distinctive motion patterns that define actions. This Relational Motion Modeling module effectively captures both individual object movements and their meaningful interactions (Figure 1(b)).

Overall, we introduce Trokens, a novel framework that leverages semantic-aware sampling and explicit motion modeling to effectively bridge motion and appearance information. Our contributions are as follows:

- We propose a novel semantic-aware point sampling approach leveraging semantic priors to adaptively distribute tracking points based on object scale and relevance.
- We develop a novel Relational Motion Modeling module that explicitly captures both intra- and inter-trajectory dynamics to understand complex motion patterns.
- We conduct comprehensive experiments with Trokens on six few-shot action recognition benchmarks including Something-Something (Full & Small) [18], Kinetics [9], UCF101 [45], HMDB51 [27], and FineGym [42], and achieve state-of-the-art performance.

2. Related Work

Few-Shot Action Recognition. Human Action Recognition requires one to model the complex temporal dynamics of a scene while also filtering out the redundant information shared between frames [26, 39, 57, 59]. While these challenges are typically addressed using large training datasets, in the setting of few-shot action recognition, methods must instead use well-constructed mechanisms to achieve effective performance with limited data. Many methods are based on metric-learning [4, 8, 15, 34, 37, 48, 58, 60, 66, 74, 75, 77], introducing various mechanisms to determine if two videos are similar or different, thus enabling action classification. Meanwhile, other methods focus on improving feature representations for spatio-temporal modeling [28, 47, 60, 63, 69, 71, 72, 79, 80]. Some recent works also leverage multi-modal language-image pretraining [40] and/or additional text data at training or inference time to further enhance performance [6, 7, 12, 19, 31, 46, 65, 67, 68]. While these works show strong results, they represent a bifurcation of the field into two domains: multi-modal and vision-only few-shot action recognition. Our method, Trokens, is a vision-only method, and we focus on primarily comparing with like baselines. A recent work [22] proposes a state space based architecture for long sequence few shot action recognition. While promising, improving architecture is orthogonal to our contributions.

Point Tracking for Feature Learning. Point tracking has a long history in computer vision research, and now recent advances in the field have enabled the efficient generation of dense and high-quality point tracks [13, 14, 21, 25, 32, 49, 56, 78]. These tracks lend themselves well to a fundamental element of video learning: the disentanglement of motion and appearance information. Several prior works have successfully applied point tracks to guide the extraction of deep and classical features [50, 51, 53]. The recent work TATs [28] further demonstrates the power of point tracking for transformer token pooling in few-shot action recognition. However, TATs falls short on our two key challenges: its uniform grid-based sampling fails to adapt to object scales, and it treats trajectories merely as feature anchors, neglecting the rich motion patterns they contain. These limitations motivate our Trokens approach that addresses both sampling and motion modeling challenges.

Motion Features. In recent years, many architectures and approaches have been proposed to learn joint or disentangled appearance and motion features from video [29, 52, 54, 73, 76]. However, such approaches are reliant on training data and struggle in low-data few-shot regimes. Meanwhile, other methods have been proposed to model motion features directly from optical flow [30, 38], or point trajectories [1, 51, 55]. In this work, we aim to improve few-shot action recognition performance by efficiently leveraging motion features derived from our trajectories. We draw inspiration from classical vision methods like Histogram of Oriented Gradients (HoG) [11], and Histogram of Oriented Displacements (HoD) [17]. Specifically, we present a new implementation of HoD tailored to toward general object intra-track motion features, which we describe in Section 4.3. While the original HoD is focused on only human skeleton keypoints, our version is designed for general object motion characterization. Additionally, we preserve temporal order through per-timestep computation rather than whole-trajectory pyramidal aggregation like [17].

3. Few-shot Action Recognition Setup

Few-shot action recognition is the problem of recognizing novel action classes with few labeled instances per class. Unlike fully supervised learning, training and test classes are mutually exclusive in the few-shot setting. Formally, given a training set $D_{\text{train}} = \{(v_i, y_i) \mid y_i \in C_{\text{train}}\}$ and test set $D_{\text{test}} = \{(v_i, y_i) \mid y_i \in C_{\text{test}}\}$, $C_{\text{train}} \cap C_{\text{test}} = \phi$. Training is done with episode-based meta-learning where each episode has a support set S with N classes and K examples per class (termed N-way K-shot, e.g., 5-way 1-shot), and a query set Q with samples to classify into these N classes.

4. Our Approach

4.1. Overview

Our work aims to leverage semantic and motion priors for few-shot video action classification. We propose the following key components: (1) Semantic-aware point trajectories sampling, where we leverage DINO features to sample semantically meaningful points for motion tracking (Sec 4.2); and (2) a Relational Motion Modeling module that explicitly models motion changes within individual trajectories and across trajectories to capture detailed movement patterns (Sec 4.3). The appearance features and the proposed motion features are sampled using the semantic-aware point trajectories and subsequently fused to form the trajectory-aligned tokens following [28]. Next, we adopt a Decoupled Space-Time Transformer [28], to process the trajectory-aligned tokens. All components of our method are trained in an end-to-end fashion using a standard few-shot loss [63].

4.2. Semantic-aware Point Trajectories

Dense point tracking [24, 25] offers promising video understanding capabilities, but its effectiveness depends on the initial selection of tracking points. The standard practice for point-based few-shot action classification is to sample points in a uniform grid [28]. Employing uniform grid sampling, despite its simplicity, often under-samples small objects crucial for understanding actions while capturing redundant information from background regions.

To address this limitation, we leverage a semantic prior to guide point selection. DINO's self-supervised learning framework produces patch tokens with rich semantic information, where tokens from the same object naturally cluster in feature space [20, 41, 43, 61, 62]. Leveraging this property, we construct semantic-aware clustering masks over patch tokens and sample points accordingly. Our strategy enables semantic-aware point sampling that adapts to object scale and semantic relevance.

Formally, we extract DINO appearance features $\mathcal{F}^{\text{RGB}} \in \mathcal{R}^{H\times W\times T\times C}$ from the input video, where T is the number of frames and H, W are spatial dimensions. We cluster the appearance features [5] into L groups which are subsequently used to sample semantic-aware points. For M trajectories, we sample $q = \frac{M}{L}$ points per cluster, from the first frame, as the semantic-aware points. We denote these points for all clusters as $P_s = \{(x_s^i, y_s^i)\}_{i=1}^{M}$, where (x_i, y_i) is the spatial coordinate of a point. These points serve as initialization for trajectory extraction.

To track the sampled points, we utilize pretrained dense point tracking model Co-tracker [25] (denoted as $\mathcal{T}$). The extracted point trajectories, known as semantic-aware point trajectories, are given by $\mathcal{P} = \mathcal{T}(P_s) = \{\mathcal{P}^m\}_{m=1}^{M}$. Each trajectory $\mathcal{P}^m = [(x_t^m, y_t^m)]_{t=1}^{T} \in \mathcal{R}^{T\times 2}$, captures the motion of a point over T frames. As shown in Fig. 1, our approach provides better coverage of small but significant objects while reducing redundant background trajectories compared to uniform sampling.

4.3. Relational Motion Modeling

Given semantic-aware point trajectories, a natural question is how to best capture their rich motion dynamics. To this end, we propose to explicitly model motion in two key aspects: dynamics within individual trajectories (intra-motion) and relationships across different trajectories (inter-motion). Such fine-grained representations capture both local motion patterns and cross trajectory interactions, providing discriminative features crucial for action recognition.

Intra-motion Module. Inspired by the Histogram of Oriented Gradients (HoG) [11], we revisit *Histogram of Displacement (HoD)* [16] to encode both magnitude and orientation changes over time on top of point trajectories.

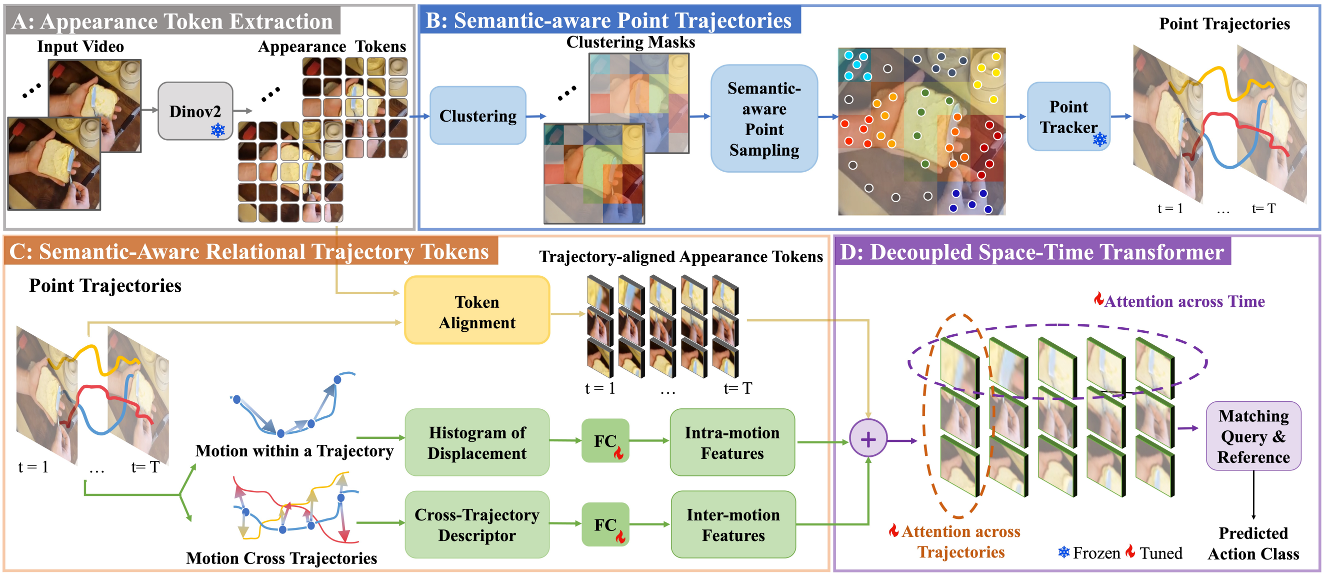

Figure 2. **Method Overview.** (A) Given an input video, we extract appearance tokens using DINOv2. (B) We then cluster these tokens and sample semantic-aware points in the initial frame, which are tracked using Co-tracker [25] to obtain point trajectories. (C) We compute intra- and inter-motion features, reorder appearance tokens via token alignment [28], and fuse them with motion features via element-wise addition to form semantic-aware relational trajectory tokens. (D) Finally, we input these tokens into a Decoupled Space-Time Transformer for few-shot action classification.

Given a trajectory $\mathcal{P}^m$, we compute the displacement at time t as $\Delta x_t = (x_t - x_{t-\delta})$ and $\Delta y_t = (y_t - y_{t-\delta})$, where δ is a hyperparameter controlling the temporal interval (we omit m for simplicity). The displacement magnitude is $\Delta d_t = \sqrt{\Delta x_t^2 + \Delta y_t^2}$, and the direction is $\theta_t = \arctan 2(\Delta y_t, \Delta x_t)$, with zero padding for $t < \delta$. For each timestep, we bin the orientation θ_t into a histogram with B bins spanning 360 degrees (*e.g.,* $B = 32$). Each displacement Δd_t contributes to the two nearest orientation bins proportionally, weighted by its magnitude. This produces a histogram of displacement descriptor for each trajectory:

$$\mathbf{H}_{\text{HoD}} = f_{\text{HoD}}(\mathcal{P}^m) \in \mathbb{R}^{T \times B}$$

We then apply a fully connected layer to project this descriptor into a C-dimensional space for all M trajectories to obtain our intra-motion features $\mathcal{F}^{\text{motion}}_{\text{intra}}$ as below:

$$\mathcal{F}^{\text{motion}}_{\text{intra}} = \text{FC}(f_{\text{HoD}}(\mathcal{P})) \in \mathbb{R}^{M \times T \times C}.$$

Our approach differs from prior work [16] in encoding HoD in several aspects. First, we preserve temporal order through per-timestep computation rather than whole-trajectory aggregation. Second, we enhance expressiveness via learnable projections. Moreover, our formulation generalize beyond human keypoints to arbitrary trajectories for broader applicability to motion analysis tasks.

Inter-motion Module. While intra-motion features capture dynamics within individual trajectories, complex actions involve coordinated movements (e.g., relative hand positions distinguish "opening a door" from "chopping vegetables"). We complement our intra-motion features with inter-motion modeling that captures evolving spatial relationships among trajectories. Specifically, we compute pairwise relative displacements between trajectories. For each trajectory $\mathcal{P}^m = [(x_t^m, y_t^m)]_{t=1}^T$ at time t, we define its cross-trajectory descriptor as:

$$\mathbf{d}_t^m = \left[(x_t^m - x_t^{m'}, y_t^m - y_t^{m'}) \right]_{m'=1}^{M} \in \mathbb{R}^{2M}$$

This captures relative positions between trajectory m and all other trajectories in a fixed order. The complete cross-trajectory descriptor $\mathbf{d} \in \mathbb{R}^{M \times T \times 2M}$ represents spatial relationships across all trajectories and timesteps. Finally, we obtain inter-motion features by projecting the cross-trajectory descriptors to the feature space:

$$\mathcal{F}^{\text{motion}}_{\text{inter}} = \text{FC}(\mathbf{d}) \in \mathbb{R}^{M \times T \times C},$$

It is worth noting that while transformers could potentially learn motion patterns through self-attention, their position embeddings primarily encode static locations without directly capturing temporal displacements or cross-trajectory relationships. Our explicit modeling of motion dynamics provides prior knowledge that helps the model focus on discriminative motion features rather than relying on self-attention to implicitly discover these patterns.

4.4. Motion-aware Space-Time Transformer

Given point trajectories $\mathcal{P}$, intra, inter motion features $\mathcal{F}_{\text{intra}}^{\text{motion}}$, $\mathcal{F}_{\text{inter}}^{\text{motion}}$, and appearance tokens from DINO $\mathcal{F}^{\text{RGB}}$, we utilize a transformer for spatiotemporal modeling of both motion and appearance. We first construct trajectory-aligned appearance tokens from point trajectories and appearance features following [28]. Given video appearance tokens $\mathcal{F}^{\text{RGB}} \in \mathbb{R}^{H\times W\times T\times C}$ and point trajectories $\mathcal{P} \in \mathbb{R}^{M\times T\times 2}$, we extract trajectory-aligned appearance tokens:

$$\mathcal{F}_{\text{traj}}^{\text{RGB}} = \text{Align}(\mathcal{F}^{\text{RGB}}, \mathcal{P}) \in \mathbb{R}^{M\times T\times C}, \tag{1}$$

where $\text{Align}(\cdot)$ samples appearance features given point coordinates. This reordering aligns visual information with motion paths, helping self-attention learn motion explicitly. We then fuse our intra-inter motion features and trajectory-aligned appearance tokens via element-wise addition, enabling the transformer to capture both intra- and inter-trajectory dependencies:

$$\mathcal{F}^{\text{fuse}} = \mathcal{F}_{\text{traj}}^{\text{RGB}} + \mathcal{F}_{\text{intra}}^{\text{motion}} + \mathcal{F}_{\text{inter}}^{\text{motion}}. \tag{2}$$

Given the semantic-aware relational trajectory tokens $\mathcal{F}^{\text{fuse}}$, we employ a decoupled attention strategy that processes the temporal and spatial dimensions separately following [28]. Self-attention is applied within each trajectory to model temporal dependencies, and across trajectories to capture spatial relationships in parallel, the results of which are then added together to form the final output embeddings $\mathcal{F}_{\text{final}}$. Finally, we conduct cross-attention between a learnable `CLS` token and the final output embeddings $\mathcal{F}_{\text{final}} \in \mathbb{R}^{M\times T\times C}$, and produced the output class token $\mathbf{c}_{\text{cls}} \in \mathcal{R}^{C}$.

4.5. Few-shot Loss

We extract the output embeddings $\mathcal{F}_{\text{final}}$ and class token $\mathbf{c}_{\text{cls}}$ for all samples in the support and query sets. To obtain the classification output, we apply a fully connected layer on the output class token, mapping it to the number of classes C_{train} in the training set with $p_{\text{cls}} = \text{FC}(\mathbf{c}_{\text{cls}}) \in \mathbb{R}^{C_{\text{train}}}$.

Following prior works [28, 60, 63, 66], our loss consists of two parts. One is a cross-entropy loss applied to the classification output on the query set p_{cls}^{Q}, capturing global C_{train} class information. The other is a contrastive loss [63] applied to final embeddings between the query set $\mathcal{F}_{\text{final}}^{\text{Q}}$ and the reference set $\mathcal{F}_{\text{final}}^{\text{S}}$ for N-way few-shot classification to encourage feature discrimination:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(p_{\text{cls}}^{Q}, y) + \mathcal{L}_{\text{Contrastive}}(\mathcal{F}_{\text{final}}^{Q}, \mathcal{F}_{\text{final}}^{S}). \tag{3}$$

We refer readers to [28] and [63] for further details.

5. Experiments

5.1. Datasets

We evaluate our approach's effectiveness across multiple action recognition benchmarks using established few-shot splits: Something-Something [18], Kinetics [9], UCF101 [45], HMDB51 [27], and FineGym [42]. For Something-Something, we use two standard configurations [8]: SSV2 Small (100 samples per class; 100 classes) and SSV2 Full (all classes). Our evaluations follow the split protocols from previous works: Kinetics splits from [79], UCF101 and HMDB51 splits from [69, 74], and FineGym splits from [28]. To ensure fair comparison, we maintain consistency with the evaluation protocols used in prior works [28, 60, 63, 66, 72].

5.2. Implementation details

We follow prior work [28, 63] for most architectural choices and training configurations. Using DINOv2-base [35], we get semantic clusters from which 256 semantic-aware points are sampled for tracking via CoTracker [24, 25]. The architecture employs a single transformer block and uses 32-bin Histogram of Directions (HoD) in the intra-motion module. During training, only the transformer and motion modules are optimized while other components remain frozen. Following standard protocols [28, 60, 63, 72], we evaluate using average few-shot accuracy across 10,000 episodes.

5.3. Quantitative Results

We evaluate Trokens against previous state-of-the-art approaches under the standard 5-way K-shot setting. Tables 1, 2, and 5 present our results on SSV2 Full and Kinetics (K=1-5), SSV2 Small/UCF-101/HMDB-51 (K=1,3,5), and FineGym respectively. On SSV2 Full, Trokens consistently outperforms TATs [28] with gains of 3.8%, 2.8%, 3.2%, 5.3%, and 2.1% across 1-5 shots respectively. For Kinetics, we achieve improvements of 1.0% and 1.2% in 1-shot and 2-shot settings, with comparable performance in higher shots. The modest gains reflect Kinetics' inherent appearance bias, where actions are primarily distinguishable through static cues, reducing the effectiveness of our motion-focused contributions. SSV2 Small demonstrates significant improvements with gains of 3.5%, 5.3%, and 4.5% across 1,3,5-shot settings. UCF-101 shows consistent improvements of 2.0%, 0.5%, and 2.4%, while HMDB-51 exhibits substantial gains of 9.8%, 8.2%, and 5.3% across respective shots. FineGym follows this trend with improvements of 2.6%, 2.3%, and 2.0% for 1,3,5-shot settings. Furthermore, when varying N-way settings (Table 3), Trokens maintains its superior performance with consistent gains of 3-4% on SSV2 Full and 1-2% on Kinetics in the 1-shot setting. These substantial improvements across diverse datasets, shot settings, and N-way configurations demonstrate our approach's robust and superior nature.

Class-wise performance anlaysis. Figure 3 presents a class-wise performance comparison between our method

Table 1. Comparison of few-shot action accuracy (1-5 shots) on SSV2 Full and Kinetics datasets versus contemporary methods. Best results are bolded, second-best underlined, and "-" indicates unavailable data.

		SSV2 Full					**Kinetics**				
Method	**Reference**	1-shot	2-shot	3-shot	4-shot	5-shot	1-shot	2-shot	3-shot	4-shot	5-shot
OTAM [8]	CVPR'20	42.8	49.1	51.5	52.0	52.3	72.2	75.9	78.7	81.9	84.2
TRX [37]	CVPR'21	42.0	53.1	57.6	61.1	64.6	63.6	76.2	81.8	83.4	85.2
STRM [47]	CVPR'22	43.1	53.3	59.1	61.7	68.1	62.9	76.4	81.1	83.8	86.7
MTFAN [69]	CVPR'22	45.7	-	-	-	60.4	74.6	-	-	-	87.4
HYRSM [60]	CVPR'22	54.3	62.2	65.1	67.9	69.0	73.7	80.0	83.5	84.6	86.1
HCL [77]	ECCV'22	47.3	54.5	59.0	62.4	64.9	73.7	79.1	82.4	84.0	85.8
Nguyen et al [33]	ECCV'22	43.8	-	-	-	61.1	74.3	-	-	-	87.4
Huang et al [23]	ECCV'22	49.3	-	-	-	66.7	73.3	-	-	-	86.4
MoLo [63]	CVPR'23	56.6	62.3	67.0	68.5	70.6	74.0	80.4	83.7	84.7	85.6
SloshNet [71]	AAA1'23	46.5	-	-	-	68.3	70.4	-	-	-	87.0
GgHM [72]	ICCV'23	54.5	-	-	-	69.2	74.9	-	-	-	87.4
RFPL [70]	ICCV'23	47.0	54.6	58.3	60.3	61.0	74.6	80.0	82.1	84.1	86.8
CCLN [64]	PAMI'24	46.0	-	-	-	61.3	75.8	82.1	85.0	86.1	87.5
HYRSM++ [66]	PR'24	55.0	63.5	66.0	68.8	69.8	74.0	80.8	83.9	85.3	86.4
TATS [28]	ECCV'24	57.7	67.1	70.0	70.6	74.6	81.9	86.5	**89.9**	90.6	91.1
Tokens	-	**61.5**	**69.9**	**73.8**	**75.9**	**76.7**	**82.9**	**87.7**	**89.9**	**90.8**	**91.2**

and previous approaches [28, 63] on both SSV2 Small and SSV2 Full splits. Our method demonstrates consistent improvements across classes through the combined benefits of semantic-aware sampling and motion modules. The semantic sampling strategy ensures better tracking of smaller objects, while our motion modules capture complex temporal dynamics effectively. On SSV2 Small, classes like 'Unfolding something' and 'Twisting something' show notable improvements, particularly benefiting from our motion-aware architecture. Similarly, SSV2 Full exhibits enhanced performance in classes such as 'Pulling something from left to right' and 'Dropping something next to something'. However, our analysis also reveals limitations in handling rapid motions causing blur (e.g., 'Rolling something on flat surface') and significant camera movements (e.g., 'Picking something up'), where point tracking becomes challenging.

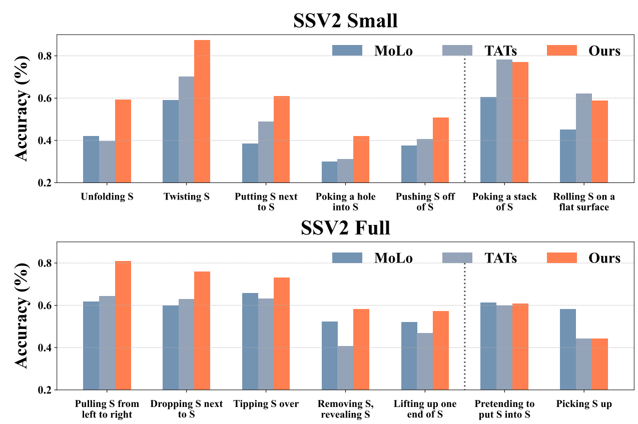

Figure 3. Class-wise accuracy comparison between MoLo [63], TATs [28] and our method. Left: classes where our approach shows performance gains. Right: classes without improvement.

5.4. Ablation Analysis

We conduct an ablation study to evaluate key design decisions. Through systematic experimentation, we analyze the impact of each component on model performance and provide empirical justification for our final configuration. We also show additional ablations in our supplementary.

Impact of each component. Table 4 presents an ablation study of our key components. Starting from the baseline configuration [28], we first evaluate our semantic-aware point sampling strategy, which yields improvements of 2% and 1.3% on SSV2 Small (1-shot and 5-shot), and 0.9% on SSV2 Full (1-shot). Given these gains, we retain this sampling strategy for subsequent experiments. We then analyze our motion modules independently. The intra-motion module alone achieves gains of $\sim$2% on both shots of SSV2 Small and up to 3.3% on SSV2 Full. Similarly, the inter-motion module independently shows improvements of 1.3–1.6% on SSV2 Small and up to 2% on SSV2 Full. When combined, these modules yield consistent improvements of 1–2% across all settings, suggesting they capture complementary motion information that enhances performance.

Analysis of Intra-motion module. Table 6 compares our choice of Histogram of Oriented Displacement (HoD) features against displacement-only features for intra-motion representation. To isolate the impact of feature choice,

Table 2. Comparison of few-shot action accuracy (1,3 and 5 shots) on SSV2 Small, UCF-101 and HMDB-51 datasets versus contemporary methods. Best results are bolded, second-best underlined, and "-" indicates unavailable data.

		SSV2 Small			UCF-101			HMDB-51		
Method	**Reference**	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
OTAM [8]	CVPR'20	36.4	45.9	48.0	79.9	87.0	88.9	54.5	65.7	68.0
TRX [37]	CVPR'21	36.0	51.9	56.7	78.2	92.4	96.1	53.1	66.8	75.6
STRM [47]	CVPR'22	37.1	49.2	55.3	80.5	92.7	96.9	52.3	67.4	77.3
MTFAN [69]	CVPR'22	-	-	-	84.8	-	95.1	-	-	-
HYRSM [60]	CVPR'22	40.6	52.3	56.1	83.9	93.0	94.7	60.3	71.7	76.0
HCL [77]	ECCV'22	38.7	49.1	55.4	82.5	91.0	93.9	59.1	71.2	76.3
Nguyen et al [33]	ECCV'22	-	-	-	-	-	-	59.6	-	76.9
Huang et al [23]	ECCV'22	38.9	-	61.6	71.4	-	91.0	60.1	-	77.0
MoLo [63]	CVPR'23	42.7	52.9	56.4	86.0	93.5	95.5	60.8	72.0	77.4
GgHM [72]	ICCV'23	-	-	-	85.2	-	96.3	61.2	-	76.9
RFPL [70]	ICCV'23	-	-	-	82.5	94.1	96.3	-	-	-
CCLN [64]	PAMI'24	-	-	-	86.9	94.2	96.1	65.1	76.2	78.8
HYRSM++ [66]	PR'24	42.8	52.4	58.0	85.8	93.5	95.9	61.5	72.7	76.4
TATs [28]	ECCV'24	47.9	60.0	64.4	92.0	96.8	95.5	60.0	71.8	77.0
Trokens	-	**53.4**	**65.3**	**68.9**	**94.0**	**97.3**	**97.9**	**69.8**	**80.0**	**82.3**

Table 3. Comparative N-way 1-shot classification accuracy (N=5-10) on Kinetics and SSV2 Full datasets versus contemporary methods. Best and second-best results are bolded and underlined, respectively.

	SSV2 Full						Kinetics					
Method	5-way	6-way	7-way	8-way	9-way	10-way	5-way	6-way	7-way	8-way	9-way	10-way
OTAM [8]	42.8	38.6	35.1	32.3	30.0	28.2	72.2	68.7	66.0	63.0	61.9	59.0
TRX [37]	42.0	41.5	36.1	33.6	32.0	30.3	63.6	59.4	56.7	54.6	53.2	51.1
HyRSM [60]	54.3	50.1	45.8	44.3	42.1	40.0	73.7	69.5	66.6	65.5	63.4	61.0
MoLo [63]	56.6	51.6	48.1	44.8	42.5	40.3	74.0	69.7	67.4	65.8	63.5	61.3
TATs [28]	57.7	55.7	52.5	50.0	47.0	45.8	81.9	79.0	76.1	75.2	72.2	72.0
Trokens	**61.5**	**59.1**	**56.5**	**54.6**	**51.4**	**49.1**	**82.9**	**80.2**	**78.5**	**76.8**	**75.5**	**73.3**

Table 4. Impact of each component on Trokens, demonstrating the relative contribution of individual elements with the final row representing our final setting.

Semantic-aware sampling	**Intra motion**	**Inter motion**	**SSV2 Small**		**SSV2 Full**	
			1-shot	5-shot	1-shot	5-shot
✗	✗	✗	47.9	64.4	57.7	74.6
✓	✗	✗	49.9	65.7	58.6	74.2
✓	✓	✗	51.8	67.7	61.3	75.7
✓	✗	✓	52.2	67.3	60.6	74.8
✓	✓	✓	**53.4**	**68.9**	**61.5**	**76.7**

we conduct all experiments with the inter-motion module disabled. In this controlled setting, HoD features consistently outperform displacement-only features on SSV2 Full, yielding gains of 2.4%, 2.0%, and 1.5% for 1-shot, 3-shot, and 5-shot settings, respectively. These results demonstrate that incorporating directional information through HoD features captures richer trajectory patterns compared to using displacement information alone.

Varying number of points. To evaluate the effectiveness of our semantic sampling strategy, we analyze performance across different point counts in the 1-shot setting on both SSV2 Small and SSV2 Full splits. Fig 4 compares our method against TATs [28] using equivalent point counts. Notably, our approach with just 32 points outperforms TATs using 256 points, demonstrating that our semantic-aware sampling yields more informative points for the task. This highlights the quality of our sampling strategy in capturing motion information with significantly fewer points.

Effect of trainable parameters. To address potential concerns about performance gains stemming from increased model capacity, we implemented a modified version of [28] with expanded parameters to match our model size.

Table 5. Comparison of few-shot action accuracy (1,3 and 5 shots) on the FineGym dataset.

	FineGym		
Method	1-shot	3-shot	5-shot
MoLo [63]	73.3	80.2	84.8
TATs [28]	81.8	86.0	87.9
Trokens	**84.4**	**88.3**	**89.9**

Table 6. Comparative analysis of intra-motion module variants independent of inter-motion module.

	SSV2 Full		
Intra-motion module	1-shot	3-shot	5-shot
Displacement only	59.9	71.6	74.2
HoD (ours)	**61.3**	**73.6**	**75.7**

Table 7. Trainable parameter analysis. * represents our implementation of [28] with increased parameters for fair comparison.

		SSV2 Small		SSV2 Full	
Method	**Params**	1-shot	5-shot	1-shot	5-shot
TATs [28]	11.8 M	47.9	64.4	57.7	74.6
TATs* [28]	18.1 M	48.0	63.0	59.6	74.3
Trokens	17.4 M	**53.4**	**68.9**	**61.5**	**76.7**

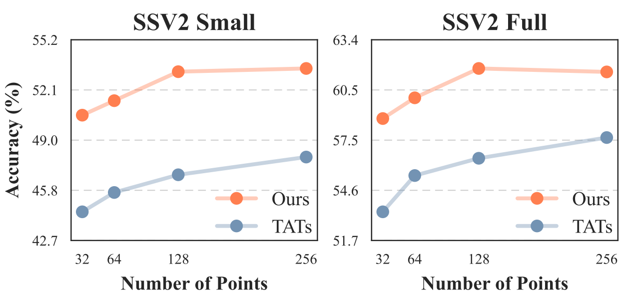

Figure 4. Few-shot accuracy by variyng the number of points on SSV2 Small and SSV2 Full for 1-shot setting.

We maintaining the rest of their original configuration. Table 7 shows that this parameter-matched re-implementation of [28] does not yield improved performance. This demonstrates that our performance gains stem from the effectiveness of our motion modules in capturing complementary information, rather than from increased model capacity.

5.5. Qualitative Analysis

Figure 5 demonstrates our semantic-aware point sampling through trajectory visualization. Our method concentrates tracking points on action-relevant objects for meaningful motion capture. The top-right quadrant showing ”Taking something out of something” tracks both a lemon and bottle, capturing the essential lifting motion. Trajectories across different examples of the same action show striking similarities while remaining distinct from other actions. This intra-class similarity and inter-class variation shows how our semantic sampling enhances action recognition by focusing on meaningful motion patterns.

6. Limitations and Future Work

While our approach demonstrates strong performance overall, it faces certain limitations. On datasets like Kinetics our motion-focused approach provides limited performance gains as its actions are distinguishable from appearance alone. Additionally, our method faces challenges inherent to point tracking: it becomes vulnerable to rapid motions that cause motion blur. Significant camera movement can disrupt point trajectory consistency, impacting performance in classes involving substantial viewpoint changes. These

Figure 5. Visualization of action trajectory similarities across four classes, where semantic-based sampling enables object-focused trajectories. Each quadrant demonstrates intra-class motion consistency while maintaining inter-class discriminative features.

limitations primarily stem from the fundamental challenges in point tracking rather than our architectural choices, suggesting potential future directions in developing more robust tracking mechanisms. While we demonstrate the effectiveness of our approach in few-shot action recognition, we hope Trokens inspires future research in both full classification settings and broader video understanding tasks, where our motion-aware architecture could prove beneficial.

7. Conclusion

In this paper, we introduced Trokens, a novel approach that advances few-shot action recognition through semantic-aware motion trajectory tokens. Our approach addresses two fundamental challenges in point tracking-based video understanding through semantic-aware sampling and a relational motion modeling framework that captures both intra-trajectory dynamics and inter-trajectory relationships. The effectiveness of our method is demonstrated through state-of-the-art performance across six challenging few-shot action recognition benchmarks. These results validate our core insight that combining semantic guidance with explicit motion modeling through trajectory tokens provides a robust foundation for understanding human actions in limited-data scenarios while opening promising directions for future research in dynamic scene understanding and broader video recognition tasks.

References

[1] Haiam A Abdul-Azim and Elsayed E Hemayed. Human action recognition using trajectory-based representation. *Egyptian Informatics Journal*, 16(2):187–198, 2015. 3

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 2

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 2

[4] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. *arXiv preprint arXiv:1907.09021*, 2019. 2

[5] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 3

[6] Congqi Cao, Yueran Zhang, Qinyi Lv, Lingtong Min, and Yanning Zhang. Exploring the adaptation strategy of clip for few-shot action recognition. In *Proceedings of the 1st International Workshop on Efficient Multimedia Computing under Limited*, pages 39–48, 2024. 2

[7] 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

[8] 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. 2, 5, 6, 7

[9] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 2, 5

[10] Seokju Cho, Jiahui Huang, Jisu Nam, Honggyu An, Seungryong Kim, and Joon-Young Lee. Local all-pair correspondence for point tracking. In *European Conference on Computer Vision*, pages 306–325. Springer, 2024. 1, 2

[11] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 886–893. Ieee, 2005. 3

[12] Fuqin Deng, Jiaming Zhong, Nannan Li, Lanhui Fu, Bingchun Jiang, Yi Ningbo, Feng Qi, He Xin, and Tin Lun Lam. Text-guided graph temporal modeling for few-shot video classification. *Engineering Applications of Artificial Intelligence*, 137:109076, 2024. 2

[13] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adrià Recasens, Lucas Smaira, Yusuf Aytar, João Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022. 2

[14] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. *arXiv preprint arXiv:2306.08637*, 2023. 2

[15] 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. 2

[16] Mohammad Abdelaziz Gowayyed, Marwan Torki, Mohamed Elsayed Hussein, and Motaz El-Saban. Histogram of oriented displacements (hod): Describing trajectories of human joints for action recognition. In *IJCAI*, pages 1351–1357, 2013. 3, 4

[17] Mohammad Abdelaziz Gowayyed, Marwan Torki, Mohammed Elsayed Hussein, and Motaz El-Saban. Histogram of oriented displacements (hod): Describing trajectories of human joints for action recognition. In *Twenty-third international joint conference on artificial intelligence*, 2013. 2, 3

[18] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter N. Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5843–5851, 2017. 2, 5

[19] Fei Guo, YiKang Wang, Han Qi, Wenping Jin, Li Zhu, and Jing Sun. Multi-view distillation based on multi-modal fusion for few-shot action recognition (clip-mdmf). *Knowledge-Based Systems*, 304:112539, 2024. 2

[20] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations*, 2022. 3

[21] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision*, pages 59–75. Springer, 2022. 2

[22] Wenbo Huang, Jinghui Zhang, Guang Li, Lei Zhang, Shuoyuan Wang, Fang Dong, Jiahui Jin, Takahiro Ogawa, and Miki Haseyama. Manta: Enhancing mamba for few-shot action recognition of long sub-sequence. *arXiv preprint arXiv:2412.07481*, 2024. 2

[23] 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. 6, 7

[24] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024. 3, 5

[25] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European Conference on Computer Vision*, pages 18–35. Springer, 2024. 1, 2, 3, 4, 5

[26] Orit Kliper-Gross, Tal Hassner, and Lior Wolf. One shot similarity metric learning for action recognition. In *Similarity-Based Pattern Recognition: First International Workshop, SIMBAD 2011, Venice, Italy, September 28-30, 2011. Proceedings 1*, pages 31–45. Springer, 2011. 2

[27] Hilde Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. Hmdb: A large video database for human motion recognition. *2011 International Conference on Computer Vision*, pages 2556–2563, 2011. 2, 5

[28] 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. 1, 2, 3, 4, 5, 6, 7, 8

[29] Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho. Motionsqueeze: Neural motion feature learning for video understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 345–362. Springer, 2020. 3

[30] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 1, 3

[31] Bozheng Li, Mushui Liu, Gaoang Wang, and Yunlong Yu. Frame order matters: A temporal sequence-aware model for few-shot action recognition. *arXiv preprint arXiv:2408.12475*, 2024. 2

[32] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. Dense optical tracking: Connecting the dots. *arXiv preprint arXiv:2312.00786*, 2023. 2

[33] Khoi D Nguyen, Quoc-Huy Tran, Khoi Nguyen, Binh-Son Hua, and Rang Nguyen. Inductive and transductive few-shot video classification via appearance and temporal alignments. In *European Conference on Computer Vision*, pages 471–487. Springer, 2022. 6, 7

[34] Xinzhe Ni, Yong Liu, Hao Wen, Yatai Ji, Jing Xiao, and Yujiu Yang. Multimodal prototype-enhanced network for few-shot action recognition. *arXiv preprint arXiv:2212.04873*, 2022. 2

[35] Maxime Oquab, Timoth'ee Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. 2, 5

[36] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems*, 34:12493–12506, 2021. 2

[37] 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. 2, 6, 7

[38] Janez Perš, Vildana Sulić, Matej Kristan, Matej Perše, Klemen Polanec, and Stanislav Kovačič. Histograms of optical flow for efficient representation of body motion. *Pattern Recognition Letters*, 31(11):1369–1376, 2010. 1, 3

[39] Ronald Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010. 2

[40] 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. 2

[41] Sai Saketh Rambhatla, Ishan Misra, Rama Chellappa, and Abhinav Shrivastava. Most: Multiple object localization with self-supervised transformers for object discovery. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15777–15788, 2023. 3

[42] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 5

[43] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. 2021. 3

[44] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 1

[45] Khurram Soomro, Amir Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *ArXiv*, abs/1212.0402, 2012. 2, 5

[46] 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*, pages 6458–6468, 2024. 2

[47] 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. 2, 6, 7

[48] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[49] Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. Dino-tracker: Taming dino for self-supervised point tracking in a single video, 2024. 2

[50] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013. 1, 2

[51] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103:60–79, 2013. 1, 2, 3

[52] Heng Wang, Du Tran, Lorenzo Torresani, and Matt Feiszli. Video modeling with correlation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 352–361, 2020. 3

[53] 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. 1, 2

[54] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1430–1439, 2018. 3

[55] Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li. Action recognition based on joint trajectory maps using convolutional neural networks. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 102–106, 2016. 3

[56] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. *arXiv preprint arXiv:2306.05422*, 2023. 2

[57] Xiang Wang, Zhiwu Qing, Ziyuan Huang, Yutong Feng, Shiwei Zhang, Jianwen Jiang, Mingqian Tang, Changxin Gao, and Nong Sang. Proposal relation network for temporal action detection. *arXiv preprint arXiv:2106.11812*, 2021. 2

[58] Xiao Wang, Weirong Ye, Zhongang Qi, Xun Zhao, Guangge Wang, Ying Shan, and Hanzi Wang. Semantic-guided relation propagation network for few-shot action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 816–825, 2021. 2

[59] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Changxin Gao, and Nong Sang. Self-supervised learning for semi-supervised temporal action proposal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1905–1914, 2021. 2

[60] 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. 2, 5, 6, 7

[61] Xudong Wang, Rohit Girdhar, Stella X. Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3124–3134, 2023. 3

[62] Xudong Wang, Ishan Misra, Ziyun Zeng, Rohit Girdhar, and Trevor Darrell. Videocutler: Surprisingly simple unsupervised video instance segmentation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22755–22764, 2023. 3

[63] 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. 2, 3, 5, 6, 7, 8

[64] 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. 6, 7

[65] 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*, 132(6):1899–1912, 2024. 2

[66] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. Hyrsm++: Hybrid relation guided temporal set matching for few-shot action recognition. *Pattern Recognition*, 147:110110, 2024. 2, 5, 6, 7

[67] Yikang Wang, Fei Guo, Li Zhu, and Yuan Guo. Sfmm: Semantic-to-frame matching with multi-classifier for few-shot action recognition. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024. 2

[68] 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, 2024. 2

[69] 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. 2, 5, 6, 7

[70] 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. 6, 7

[71] Jiazheng Xing, Mengmeng Wang, Boyu Mu, and Yong Liu. Revisiting the spatial and temporal modeling for few-shot action recognition. In *AAAI Conference on Artificial Intelligence*, 2023. 2, 6

[72] 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*, pages 1740–1750, 2023. 2, 5, 6, 7

[73] Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 5682–5692, 2023. 3

[74] 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. 2, 5

[75] Songyang Zhang, Jiale Zhou, and Xuming He. Learning implicit temporal alignment for few-shot video classification. *arXiv preprint arXiv:2105.04823*, 2021. 2

[76] Yue Zhao, Yuanjun Xiong, and Dahua Lin. Recognize actions by disentangling components of dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6566–6575, 2018. 3

[77] Sipeng Zheng, Shizhe Chen, and Qin Jin. Few-shot action recognition with hierarchical matching and contrastive learning. In *European Conference on Computer Vision*, pages 297–313. Springer, 2022. 2, 6, 7

[78] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19855–19865, 2023. 2

[79] 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. 2, 5

[80] Linchao Zhu and Yi Yang. Label independent memory for semi-supervised few-shot video classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 273–285, 2020. 2