MOVE: Motion-Guided Few-Shot Video Object Segmentation

Kaining Ying* Hengrui Hu* Henghui Ding✉
Fudan University, China
https://henghuiding.com/MOVE/

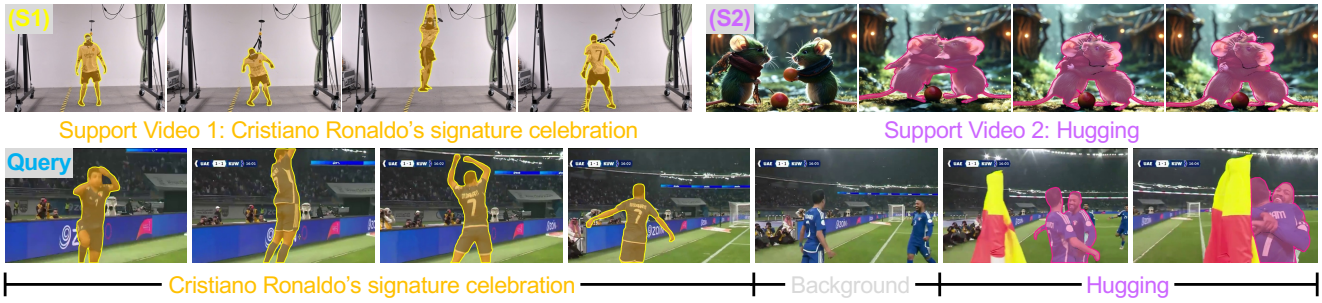


Figure 1. We propose a new benchmark for **MO**tion-guided Few-shot **V**ideo object s**E**gmentation (**MOVE**). In this example, given two support videos showing distinct motion patterns (S1: Cristiano Ronaldo's signature celebration [19], S2: hugging), our benchmark aims to segment target objects in the query video that perform the same motions as in the support videos. **MOVE** provides a platform for advancing few-shot video analysis and perception by enabling the segmentation of diverse objects that exhibit the same motions.

Abstract

*This work addresses motion-guided few-shot video object segmentation (FSVOS), which aims to segment dynamic objects in videos based on a few annotated examples with the same motion patterns. Existing FSVOS datasets and methods typically focus on object categories, which are static attributes that ignore the rich temporal dynamics in videos, limiting their application in scenarios requiring motion understanding. To fill this gap, we introduce **MOVE**, a large-scale dataset specifically designed for motion-guided FSVOS. Based on MOVE, we comprehensively evaluate 6 state-of-the-art methods from 3 different related tasks across 2 experimental settings. Our results reveal that current methods struggle to address motion-guided FSVOS, prompting us to analyze the associated challenges and propose a baseline method, Decoupled Motion-Appearance Network (**DMA**). Experiments demonstrate that our approach achieves superior performance in few-shot motion understanding, establishing a solid foundation for future research in this direction.*

*Equal contribution.
✉ Henghui Ding (henghui.ding@gmail.com) is the corresponding author with the Institute of Big Data, College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China.

1. Introduction

Few-shot video object segmentation (FSVOS) [3, 37, 40, 42, 52] aims to segment objects of unseen classes in videos using only a few annotated examples. FSVOS is a relatively underexplored yet promising field. By requiring only minimal supervision, FSVOS significantly reduces the need for extensive labeled datasets while enabling rapid adaptation to new object classes. With these advantages, it shows great potential in autonomous driving, robotics, surveillance, augmented reality, and media production [72].

Previous FSVOS methods [3, 51, 54] are semantic-centric and primarily focus on object categories, associating query videos with support sets based on object class. For example, given support images containing pandas, these methods aim to segment all pandas in the query video regardless of their individual characteristics. This semantic-centric paradigm, similar to the widely-studied few-shot image segmentation (FSS) [57, 59, 60, 69], largely overlooks the crucial temporal dynamics inherent in videos, such as object motions and temporal dependencies, thus limiting the advancement of FSVOS research.

We emphasize the fundamental role of motion patterns in videos, which cannot be adequately captured by static image-based segmentation approaches. Consider Cristiano Ronaldo's celebration motion shown in Figure 1 (S1), such

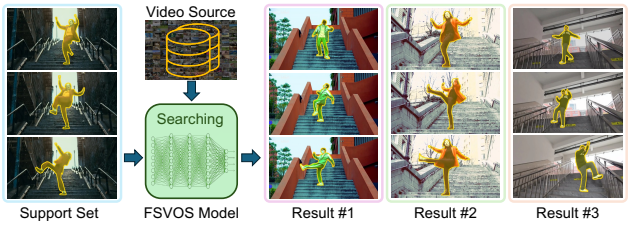


Figure 2. Given a motion of interest, our approach enables retrieval and indexing of relevant videos and their corresponding objects from the internet or personal collections. Notably, these motions of interest can be novel actions that are difficult to describe accurately using single-frame images or text alone.

dynamic patterns can only be properly represented through video sequences. Unlike previous few-shot video segmentation methods that focus on object categories, *e.g.*, *robot* or *mouse* in the support videos of Figure 1, we explore to segment objects based on their motion patterns, *e.g.*, *hugging*, regardless of their object categories. This enables us to recognize and segment objects performing the same motions, which facilitates novel motion-based object-level retrieval in video, as shown in Figure 2.

Recent referring video object segmentation (RVOS) [11, 13, 18, 30, 36, 61, 68] methods explore motion-guided expressions to segment target objects in videos. However, these methods face inherent limitations when dealing with novel or complex motions that are difficult to describe textually. In contrast, such motions can be effectively characterized by providing a reference video example, such as the distinctive dance sequence from the movie in Figure 2. While widely recognized actions eventually receive distinctive names, *e.g.*, "*CR7's celebration*" and "*Joker's dance*", this does not contradict our approach. A video is worth a thousand words, particularly for newly emerging actions that have not yet gained widespread recognition.

In light of this, we propose **MOVE**, a new large-scale motion-guided FSVOS dataset containing 224 motion categories, 4,300 videos with 261,920 frames in total, and 314,619 high-quality segmentation masks annotating 5,135 objects across 88 object categories. This dataset is designed to capture diverse motion patterns, facilitating the development and evaluation of motion-centric FSVOS methods.

By adapting existing methods [3, 11, 54] to MOVE, we find that MOVE presents great challenges in understanding and matching motions between support and query videos. Understanding motions in support videos requires comprehensive analysis of the entire video sequence, rather than relying on static single-frame semantic recognition. Furthermore, effectively extracting motion-related prototypes presents another significant challenge, as existing methods primarily focus on extracting semantic features while overlooking the dynamic information inherent in support videos. To address these challenges, we propose a Decoupled Motion-Appearance (DMA) module for extracting temporally decoupled motion-appearance prototypes, enabling the model to focus more on object motions rather than object appearance. Experiments demonstrate that our proposed DMA helps the model learn motion-centric features, thereby effectively improving model performance. We conduct a comprehensive evaluation of 6 state-of-the-art methods from 3 different related tasks across 2 experimental settings in 2 backbones, demonstrating the superiority of the proposed DMA in few-shot motion understanding.

In summary, this work makes the following three main contributions: **i)** We introduce MOVE, a motion-guided few-shot video object segmentation dataset that shifts the focus from static object categories to dynamic motion understanding. **ii)** We propose DMA, a method based on decoupled motion and appearance, which demonstrates effective few-shot motion understanding and achieves strong performance on the proposed MOVE. **iii)** We conduct comprehensive experiments, benchmarking 6 baselines on MOVE, providing a solid foundation for future research.

2. Related Work

2.1. Video Object Segmentation

Video Object Segmentation (VOS) [12, 14, 20, 21, 46] aims to track and segment the corresponding objects in a video sequence, given the object mask in the initial frame. Early deep neural network (DNN)-based methods, such as OSVOS [2] and MoNet [58], fine-tuned network parameters during inference to model inter-frame correlations. Methods like OSMN [62] and LML [1] used the first frame with a mask as a prompt to generate a prototype for pixel-level matching with subsequent frames. Recent trends have shifted toward memory-based methods. STM [43] introduced memory modules to store historical frame information, while STCN [7] enhanced memory usage efficiency. XMem [6] and Cutie [8] further improved memory mechanisms with multiple granularities and object-specific storage. Recently, SAM2 [47] emerged as a large video model extending SAM [26], achieving significant performance improvements. While previous VOS methods have achieved milestone progress, their feasibility and scalability are still constrained by the need for large volumes of densely annotated masks. Additionally, many methods struggle with out-of-domain inputs. In contrast to conventional VOS settings, we focus on few-shot settings to significantly reduce annotation costs and improve generalization to a wider range of scenarios.

2.2. Few-Shot Video Object Segmentation

Few-shot video object segmentation (FSVOS) [3, 37, 40, 42, 52] has emerged as a promising solution to address the heavy dependency on pixel-wise annotations in traditional

Table 1. Comparison of related few-shot video object segmentation/detection datasets [3, 16, 51] with our proposed dataset MOVE.

Dataset	Venue	Label Type	Annotation	Support Type	Categories	Videos	Objects	Frames	Masks
FSVOD-500 [16]	[ECCV'22]	Object	Box	Image	**500**	4,272	4,663	96,609	104,495
YouTube-VIS [3, 63]	[CVPR'21]	Object	Mask	Image	40	2,238	3,774	61,845	97,110
MiniVSPW [51]	[IJCV'25]	Object	Mask	Image	20	2,471	-	**541,007**	-
MOVE (ours)	[ICCV'25]	Motion	Mask	Video	224	**4,300**	**5,135**	261,920	**314,619**

VOS. Given an annotated support set, FSVOS aims to segment novel object categories unseen during training in query videos with only a few prompt images and masks. Recent FSVOS methods mainly focus on prototype learning [3, 37, 54] or affinity calculation [52]. DANet [3] first defined FSVOS and proposed sampled query agents for attention. TTI [52] and VIPMT [37] focused on temporal consistency through prototypes at different granularities. HPAN [41] and CoCoNet [40] further improved temporal modeling with graph attention and optimal transport respectively. While these studies advance few-shot segmentation for novel object categories, they are inherently category-centric, limiting their real-world applicability. In contrast, we propose a motion-centric approach that prioritizes the motion over the object category. In our new task, MOVE, the support set consists of videos and masks specifying a particular motion, and the model segments objects performing the same motion in query videos regardless of their categories, enabling generalization to both novel motions and objects.

2.3. Motion-centric Tasks

Motion understanding has evolved as a core research direction in video analysis, progressing from early human-centered action recognition [49, 56, 73] to more complex tasks including action detection [50, 64, 71], spatio-temporal localization [5, 28, 44]. Recent LVLMs [22, 38, 67] also pay attention to temporal motion-related tasks. However, these methods require extensive training data and cannot accurately segment target objects in videos. Recently, referring video object segmentation [11, 18, 25, 30] has explored using motion-related expressions to segment target objects, but expressions often fail to accurately describe novel motions. Our proposed DMA can learn novel actions with minimal data to segment target objects in query videos, enabling broader applications across diverse real-world scenarios.

3. MOVE Benchmark

3.1. Task Setting

Revisit of FSVOS. Few-Shot Video Object Segmentation (FSVOS) [3, 51, 54] aims to learn a segmentation model that can generalize to novel object categories with limited labeled examples. The framework operates on two disjoint data splits: a base class training set $\mathcal{D}_{train}$ and a novel class test set $\mathcal{D}_{test}$. The evaluation protocol involves episodes, where each episode comprises a support set $\mathcal{S}$ and a query set $\mathcal{Q}$. Specifically, the support set encompasses K pairs of images and their corresponding masks $\{(I_k^s, M_{k,c}^s)\}_{k=1}^K$ extracted from separate videos, with I_k^s denoting the k-th support image and $M_{k,c}^s$ representing its associated mask for class c. The query set contains a video with T frames $\{(I_t^q, M_{t,c}^q)\}_{t=1}^T$, where I_t^q indicates the t-th frame and $M_{t,c}^q$ denotes its ground truth mask. The objective of FSVOS is to leverage the visual and semantic cues from support samples to accurately segment target objects in query video frames.

Extension. The proposed MOVE focuses on motion categories rather than object categories. Since static images inherently lack the capacity to represent temporal dynamics, we extend the support set $\mathcal{S}$ to contain K video-mask pairs $\{(V_k^s, M_{k,c}^s)\}_{k=1}^K$ sampled from different videos, where each video clip V_k^s demonstrates a motion pattern with corresponding mask sequence $M_{k,c}^s$ of motion-class c. The query set $\mathcal{Q}$ remains as a video sequence of T frames $\{(I_t^q, M_{t,c}^q)\}_{t=1}^T$. This formulation shifts the focus from static appearances to temporal modeling, emphasizing motion as the core feature for video understanding.

3.2. Dataset Annotation

Vocabulary Collection. Following previous video recognition datasets [9, 24, 53], we build a hierarchical vocabulary set with four areas: *daily actions*, *sports*, *entertainment activities*, and *special actions*. Each category follows three criteria: fine-grained, mutual exclusion (clear semantic boundaries), and novelty (not well covered in existing datasets). This systematic classification lays the foundation for motion-guided few-shot video segmentation tasks.

Video Clip Collection. Videos in MOVE are sourced from two parts: **i)** public action recognition datasets [4, 9, 15, 23, 27, 29, 31, 32, 45, 70] and **ii)** internet videos under a *Creative Commons License*. During the selection process, we followed these criteria: videos should have clear motion boundaries, diverse scenes, and varied subject categories.

Mask Annotation. For videos without preexisting masks, we recruited well-trained annotators to label high-quality masks with the assistance of a state-of-the-art VOS segmentation model [48] on an interactive annotation platform.

3.3. Data Statistics and Analysis

As shown in Table 1, our MOVE benchmark contains 224 action categories across four domains (daily actions, sports,

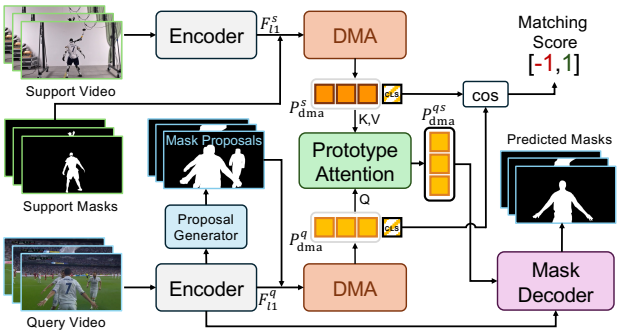


Figure 3. Overview of our proposed method.

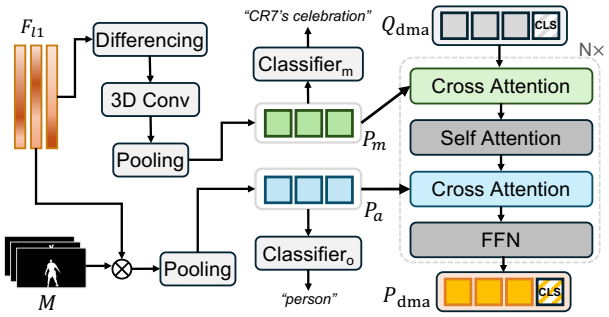


Figure 4. Decoupled Motion-Appearance (DMA) Module.

entertainment activities, and special actions), with 4,300 video clips, 5,135 moving objects, 261,920 frames, and 314,619 mask annotations. Compared to existing object-centric datasets, MOVE features video-level support samples and motion-based categories, while maintaining comparable scale in terms of videos and annotations. Each video clip is equipped with high-quality pixel-level mask annotations, capturing diverse scenes, subjects (person, vehicle, animal, *etc.*), and motion complexities. For more statistics, please refer to the supplementary materials.

4. Methodology

4.1. Overview

As shown in Figure 3, the proposed method consists of five main components: 1) a shared encoder for extracting multi-scale features from both support and query video frames, 2) proposal generator for obtaining coarse mask proposals of the query video, 3) a shared DMA module for extracting decoupled motion-appearance prototypes, 4) prototype attention module for facilitating interaction between support and query prototypes, and 5) mask decoder for generating the final segmentation masks of the query video. In the following sections, we describe each component in detail. For simplicity, we describe our method in the *1-way-1-shot* setting, although it can be easily extended to *N-way-K-shot* scenarios. Given a support video clip with T_s frames $\{I_t^s\}_{t=1}^{T_s}$ and corresponding mask sequence $\{M_t^s\}_{t=1}^{T_s}$, along with a query video clip containing T_q frames $\{I_t^q\}_{t=1}^{T_q}$, our goal is to segment out the target object mask sequence $\{\hat{M}_t^q\}_{t=1}^{T_q}$ in the query video that exhibits the same motion pattern as the object in the support video.

4.2. Encoder and Proposal Generator

Encoder. Our encoder $\mathcal{E}$ combines a backbone [17, 39] with a feature pyramid network [34] to extract multi-scale features from both the support and query videos as follows:

$$F_{l1,t}, F_{l2,t}, F_{l3,t}, F_{l4,t} = \mathcal{E}(I_t), \quad t = 1, \ldots, T, \tag{1}$$

where $F_{li,t}$ is the i-th layer feature of the t-th frame I_t. T denotes the total number of frames. $F_{l1,t}$, $F_{l2,t}$, $F_{l3,t}$, and $F_{l4,t}$ correspond to features at resolutions of $1/4$, $1/8$, $1/16$, and $1/32$ of input resolution, respectively.

Proposal Generator. This module processes multi-scale query features $\{F_{l1}^q, F_{l2}^q, F_{l3}^q, F_{l4}^q\}$ to generate coarse mask proposals. It employs three convolutional blocks at different scales ($1/32$, $1/16$, and $1/8$ resolution) with residual connections. Features are progressively refined through upsampling and fusion, with final predictions generated by a lightweight convolutional head that outputs single-channel proposals. This approach effectively balances multi-scale information utilization and computational efficiency.

4.3. Decoupled Motion-Appearance Module

As shown in Figure 4, DMA module extracts decoupled motion-appearance prototypes for both query and support branches. The module takes the $1/4$ resolution features F_{l1} and corresponding object masks M as input, where the support branch utilizes pre-annotated support masks while the query branch leverages mask proposals generated by the proposal generator.

Appearance Prototype. DMA first extracts appearance prototypes P_a by applying mask pooling on feature F_{l1}:

$$P_a = \frac{\sum_{h,w} F_{l1} \odot M}{\sum_{h,w} M} \in \mathbb{R}^{T \times d}, \tag{2}$$

where $\odot$ denotes element-wise multiplication.

Motion Prototype. DMA then extracts motion prototypes by calculating temporal differences between adjacent frames features, where the temporal difference at the last time step is padded with zeros:

$$\begin{aligned} D_{l1,t} &= F_{l1,t+1} - F_{l1,t}, \ t = 1, \ldots, T-1, \\ P_m &= \text{Pooling}(\text{Conv3D}(D_{l1})) \in \mathbb{R}^{T \times d}, \end{aligned} \tag{3}$$

where Conv3D denotes 3D convolutional layers for temporal feature enhancement, and Pooling is a spatial pooling operation that aggregates the motion features across the spatial dimensions into motion prototypes P_m.

To guide the learning of discriminative and complementary motion and appearance prototypes, we introduce two auxiliary classification heads:

$$p_o = \text{Classifier}_o(\text{AvgPool}(P_a)) \in \mathbb{R}^{C_o}, \quad (4)$$

$$p_m = \text{Classifier}_m(\text{AvgPool}(P_m)) \in \mathbb{R}^{C_m}, \quad (5)$$

where C_o is the number of predefined object categories and C_m is the number of motion categories. These classification tasks explicitly guide P_a to encode object-specific appearance information, while P_m focuses on motion-specific temporal dynamics. This decoupled supervision ensures that the two prototype branches learn complementary features for appearance and motion, respectively.

The extracted motion prototypes P_m are further refined using a transformer-based architecture. As shown in Figure 4, this architecture processes learnable queries Q_{dma} and a special [CLS] token through multiple transformer layers. Each transformer layer consists of cross-attention modules attending to motion prototypes P_m and appearance prototypes P_a, followed by self-attention modules and feed-forward networks (FFN). This process produces the final decoupled motion-appearance prototypes P_{dma} along with the [CLS] token used for prototype matching.

4.4. Prototype Attention and Mask Decoder

Prototype Attention. To fuse prototype features from both support and query videos, we introduce a prototype attention module that consists of multiple transformer layers. Given the decoupled motion-appearance prototypes P^s_{dma} and P^q_{dma}, this module performs cross-attention where P^q_{dma} serves as queries while P^s_{dma} serves as keys and values, followed by self-attention on the enhanced P^q_{dma} features. Through multiple transformer layers, this iterative process refines the prototypes, facilitating effective information exchange while preserving their distinctive characteristics. The enhanced prototypes, denoted as P^{qs}_{dma}, are subsequently used for mask generation in Mask Decoder.

Mask Decoder. The Mask Decoder generates segmentation masks by fusing multi-scale features under the guidance of prototypes P^{qs}_{dma}. It enhances features at different scales via cross-attention with prototypes, enabling the features to focus on motion-centric information. These enhanced features are then progressively fused in a top-down manner. This hierarchical design together with motion prototype guidance ensure that both high-level semantic information and low-level spatial details are effectively leveraged, contributing to the accurate prediction of the final mask.

Matching Score. To determine whether the query instance exhibits the same motion as the support instance, we compute a matching score based on the [CLS] tokens from both branches. The matching score is calculated as:

$$S_{\text{match}} = \cos([\text{CLS}]_s, [\text{CLS}]_q), \quad (6)$$

Table 2. Necessity study of the proposed MOVE benchmark.

Methods	Type	Support	Query	YTVIS [3]	MOVE
SCCAN [59]	FSS	Image	Image	62.3	40.6
HPAN [3]	FSVOS	Image	Video	63.0	44.4
HPAN*	FSVOS	Video	Video	62.7	46.3
LMPM [11]	RVOS	Text	Video	62.5	41.8

where $\cos(\cdot,\cdot)$ represents the cosine similarity between the [CLS] tokens from support and query branches. The matching score S_{match} ranges from -1 to 1, with higher values indicating that the query instance is performing the same motion as the support instance, and lower values suggesting different motions.

5. Experiment

Evaluation Metrics. Following prior works [3, 12, 54], we use $\mathcal{J}\&\mathcal{F}$ to evaluate segmentation quality, with $\mathcal{J}$ and $\mathcal{F}$ measuring IoU and contour accuracy, respectively. For robustness evaluation, we include query samples with empty foreground and adopt N-Acc and T-Acc metrics [35] to measure accuracy on empty and non-empty target samples.

Dataset Settings. MOVE contains 224 motion categories. For cross-validation, we split these into 4 folds with two strategies: **O**verlapping **S**plit (**OS**) and **N**on-overlapping **S**plit (**NS**) based on node-level motion distribution. Please refer to supplementary materials for more details.

Implementation Details. Our backbone uses ResNet50 [17] pre-trained on ImageNet [10] and VideoSwin-Tiny [39] pre-trained on Kinetics-400 [24]. Following previous work [3, 54, 65, 66], we employ both cross-entropy and IoU losses for mask prediction and proposal generation. Additionally, we use cross-entropy loss for the auxiliary classification head and matching score prediction. We use a learning rate of 1e-5 with a cosine annealing scheduler for optimization. For our main experiments, we train for 240,000 episodes on 3 folds and test on the remaining fold with 20,000 episodes, using both *2-way-1-shot* and *5-way-1-shot* settings as our primary configurations. Unless otherwise specified, ablation studies are conducted with 150,000 episodes, training on 2 folds and testing on the remaining 2 folds, using the *2-way-1-shot* setting on OS. All the experiments are conducted on 4 NVIDIA RTX A6000 (48GB) GPUs.

5.1. Benchmark Necessity Study

To demonstrate the necessity of our MOVE benchmark, we conduct experiments comparing state-of-the-art methods across different areas on both the common-used YouTube-VIS (YTVIS) [63] and our proposed MOVE datasets, as shown in Table 2. The use of YouTube-VIS strictly follows the few-shot setting in [3]. Notably, when evaluated on YTVIS, the image-based FSS method SCCAN [59] achieves 62.3% $\mathcal{J}\&\mathcal{F}$, comparable to HPAN [3] (63.0%

Table 3. Main results on MOVE benchmark with overlapping split (OS) setting. **Bold** and underlined indicate the largest and second largest values under the same backbone, respectively. VSwin-T indicates VideoSwin-T backbone [39].

Methods	Venue	Type	Backbone	Mean (2-way-1-shot)			$\mathcal{J}\&\mathcal{F}$ (2-way-1-shot)				Mean (5-way-1-shot)			$\mathcal{J}\&\mathcal{F}$ (5-way-1-shot)			
				$\mathcal{J}\&\mathcal{F}$	T-Acc	N-Acc	1	2	3	4	$\mathcal{J}\&\mathcal{F}$	T-Acc	N-Acc	1	2	3	4
LMPM [11]	[ICCV'23]	RVOS	ResNet50	41.8	93.1	5.3	45.2	42.1	40.7	39.1	26.3	98.3	2.6	27.5	31.7	22.7	23.3
CyCTR [69]	[ECCV'24]	FSS	ResNet50	34.4	98.4	1.2	32.8	34.4	35.7	34.5	22.5	99.2	0.1	23.1	21.5	20.8	24.7
SCCAN [59]	[ECCV'24]	FSS	ResNet50	40.6	93.9	5.8	47.5	37.1	40.5	37.4	28.6	97.3	2.8	27.7	32.5	27.2	27.1
DANet [3]	[CVPR'21]	FSVOS	ResNet50	45.4	97.1	8.2	41.4	44.7	47.1	48.2	25.4	77.2	28.0	27.4	23.5	25.8	25.0
HPAN [54]	[CSVT'24]	FSVOS	ResNet50	44.4	97.3	7.2	48.4	45.2	43.4	40.8	34.0	99.1	3.1	37.6	34.8	34.6	29.1
TTI [51]	[IJCV'25]	FSVOS	ResNet50	45.2	97.6	9.4	45.8	43.9	43.7	47.4	35.6	70.6	26.2	33.8	35.9	34.8	37.8
DMA (Ours)	[ICCV'25]	FSVOS	ResNet50	**50.1**	**98.6**	**11.5**	**51.2**	**46.2**	**54.3**	**48.6**	**40.2**	**99.5**	**28.7**	**40.7**	**38.9**	**41.3**	**39.7**
DANet [3]	[CVPR'21]	FSVOS	VSwin-T	49.8	93.4	16.5	49.3	47.5	52.5	49.9	36.1	37.2	30.3	34.8	34.3	38.3	37.1
DMA (Ours)	[ICCV'25]	FSVOS	VSwin-T	**51.5**	**98.9**	**21.2**	**51.1**	**48.6**	**56.3**	**50.0**	**41.4**	**99.8**	**31.0**	**41.5**	**39.8**	**42.7**	**41.1**

Table 4. Main results on MOVE benchmark with non-overlapping split (NS) setting.

Methods	Venue	Type	Backbone	Mean (2-way-1-shot)			$\mathcal{J}\&\mathcal{F}$ (2-way-1-shot)				Mean (5-way-1-shot)			$\mathcal{J}\&\mathcal{F}$ (5-way-1-shot)			
				$\mathcal{J}\&\mathcal{F}$	T-Acc	N-Acc	1	2	3	4	$\mathcal{J}\&\mathcal{F}$	T-Acc	N-Acc	1	2	3	4
LMPM [11]	[ICCV'23]	RVOS	ResNet50	38.8	94.8	4.4	45.5	34.5	36.5	38.5	29.8	96.6	2.4	28.9	26.4	37.6	26.1
CyCTR [69]	[ECCV'24]	FSS	ResNet50	28.2	98.0	1.0	31.2	25.7	33.0	22.7	23.4	95.3	3.2	23.6	21.2	27.6	21.3
SCCAN [59]	[ECCV'24]	FSS	ResNet50	34.5	92.3	5.9	41.8	32.7	29.2	34.3	27.8	96.0	4.3	31.7	31.4	25.8	22.2
DANet [3]	[CVPR'21]	FSVOS	ResNet50	44.6	97.7	2.5	48.4	36.9	49.5	43.6	29.9	96.6	4.2	29.5	24.0	31.0	35.3
HPAN [54]	[CSVT'24]	FSVOS	ResNet50	39.1	96.3	1.4	49.1	34.9	40.0	32.3	30.2	99.1	1.1	35.3	28.5	30.3	26.6
TTI [51]	[IJCV'25]	FSVOS	ResNet50	43.6	97.2	2.4	47.2	33.4	50.0	43.9	32.7	98.3	0.9	35.0	29.8	35.7	30.2
DMA (Ours)	[ICCV'25]	FSVOS	ResNet50	**46.0**	**98.2**	**7.8**	**47.8**	**37.9**	**48.0**	**50.3**	**34.7**	**99.6**	**5.0**	**35.6**	**31.5**	**37.0**	**34.6**
DANet [3]	[CVPR'21]	FSVOS	VSwin-T	47.4	97.2	1.2	53.2	37.4	48.3	50.9	30.0	74.8	2.9	34.6	26.5	32.1	26.8
DMA (Ours)	[ICCV'25]	FSVOS	VSwin-T	**49.0**	**98.0**	**8.8**	**54.4**	**37.4**	**48.5**	**55.9**	**35.4**	**97.4**	**9.3**	**37.9**	**29.9**	**36.6**	**37.2**

$\mathcal{J}\&\mathcal{F}$) which is specifically designed for FSVOS. This suggests that YTVIS primarily relies on category-based object association, requiring minimal temporal understanding between support and query samples. However, the performance landscape changes dramatically on MOVE. SCCAN's performance drops significantly to 40.6% $\mathcal{J}\&\mathcal{F}$, substantially lower than HPAN's 44.4% $\mathcal{J}\&\mathcal{F}$. This stark contrast highlights the critical role of temporal information in MOVE. Furthermore, we enhance HPAN (denoted as HPAN*) by incorporating temporal modeling during prototype extraction through a simple self-attention mechanism across frames. This modification yields a notable improvement from 44.4% to 46.3% $\mathcal{J}\&\mathcal{F}$, further emphasizing the importance of motion understanding in our benchmark.

Furthermore, we benchmark the referring video object segmentation method LMPM [11] on our MOVE dataset by converting support set into referring expressions with the template "*The one [**motion category**]*". While LMPM achieves competitive performance of 62.5% $\mathcal{J}\&\mathcal{F}$ on YTVIS, only 0.2% $\mathcal{J}\&\mathcal{F}$ lower than the temporally-enhanced HPAN*, its performance drops significantly to 41.8% $\mathcal{J}\&\mathcal{F}$ on MOVE, substantially underperforming compared to HPAN*'s 46.3% $\mathcal{J}\&\mathcal{F}$. We attribute this performance gap to the presence of fine-grained, novel, and specialized motion patterns in our MOVE dataset like mutations and moonwalks, which are difficult to describe clearly using text. These findings underscore the unique challenges posed by MOVE and emphasize its necessity in advancing motion-centric few-shot video understanding.

5.2. Main Results

As shown in Table 3 and Table 4, we benchmark referring video object segmentation (RVOS) [11], few-shot image segmentation (FSS) [59, 69], and few-shot video object segmentation (FSVOS) methods [3, 51, 54] across *2-way-1-shot* and *5-way-1-shot* test settings on both OS and NS data splits with two different backbones, ResNet50 [17] and VideoSwin-T [39]. Our proposed DMA consistently outperforms all competing methods across all metrics and settings, demonstrating its superior few-shot motion understanding and segmentation capabilities. For the $\mathcal{J}\&\mathcal{F}$ metric with ResNet50 backbone, DMA achieves significant improvements over the second-best method, reaching 50.1% (*vs*. 45.4%) in *2-way-1-shot* and 40.2% (*vs*. 35.6%) in *5-way-1-shot* under the OS setting. This substantial performance gap highlights the limitations of existing methods in effectively modeling motion patterns. When using VideoSwin-T backbone, which provides better temporal feature extraction, our method further improves to 51.5% and 41.4% in respective settings, indicating the importance of temporal modeling in motion-centric segmentation. It is worth noting that performance on the NS setting (46.0% $\mathcal{J}\&\mathcal{F}$ with ResNet50) is lower than OS (50.8%), reflecting its greater challenge as a more realistic scenario where test categories have completely different parent classes from training categories. Regarding robustness metrics, while DMA maintains high target accuracy (T-Acc) of 98.6%

Table 5. Ablation study on motion extractor.

ID	Motion Extractor	$\mathcal{J}\&\mathcal{F}$	T-acc	N-acc
I	Mask Pooling	41.3	98.0	6.8
II	Mask Adapter	43.4	98.4	6.6
III	Differencing	**46.8**	**99.8**	**12.3**

Table 6. Ablation study on DMA prototype extractor.

ID	Appear.	Motion	$\mathcal{J}\&\mathcal{F}$	T-acc	N-acc
I	✓	✗	36.5	80.1	**30.4**
II	✗	✓	43.8	95.5	10.7
III	✓	✓	**46.8**	**99.8**	12.3

and achieves better non-target accuracy (N-Acc) of 11.5% with ResNet50 compared to baselines, the generally low N-Acc scores across all methods suggest a common challenge in effectively modeling background information to reduce false positives. This limitation points to a promising direction for future research in MOVE.

5.3. Ablation Studies

Ablation study on motion extractor. As shown in Table 5, our proposed differencing-based motion extractor achieves better performance compared to baseline approaches such as mask pooling and mask adapter [33], improving $\mathcal{J}\&\mathcal{F}$ from 41.3% (mask pooling) and 43.4% (mask adapter) to 46.8%. The explicit motion modeling through frame differencing enables more effective extraction of motion-centric prototypes, leading to enhanced motion pattern recognition and segmentation performance.

Ablation study on DMA prototype extractor. As shown in Table 6, we analyze the contribution of appearance and motion prototypes in our DMA prototype extractor. Using only appearance prototypes (I) achieves 36.5% $\mathcal{J}\&\mathcal{F}$, while using only motion prototypes (II) results in 43.8% $\mathcal{J}\&\mathcal{F}$. These results suggest that while appearance features provide static cues for target object recognition, they are insufficient for motion-centric video understanding in MOVE. In contrast, motion features capture temporal dynamics, enhancing the distinction between different motion categories. When combining both prototypes (III), our model achieves the best performance of 46.8% $\mathcal{J}\&\mathcal{F}$, demonstrating the complementary nature of static appearance and dynamic motion information in our DMA mechanism.

Ablation study on auxiliary classification. As shown in Table 7, applying auxiliary classification supervision separately to appearance and motion prototypes yields the best performance of 46.8% $\mathcal{J}\&\mathcal{F}$. We attribute this improvement to explicit supervision of object and motion, which effectively enhances the decoupling of motion and appearance features, resulting in overall performance gains.

Orcale results. As shown in Table 8, we conduct oracle experiments to analyze the performance upper bound of our model. When provided with perfect motion category labels, the model achieves 63.6% $\mathcal{J}\&\mathcal{F}$, while using ground truth masks yields higher $\mathcal{J}\&\mathcal{F}$ of 74.3%. The results indicate significant room for improvement in both motion understanding and mask prediction capabilities.

Table 7. Ablation study on auxiliary classification.

ID	Object	Motion	$\mathcal{J}\&\mathcal{F}$	T-acc	N-acc
I	✗	✗	43.8	97.2	5.2
II	✗	✓	44.2	87.6	9.6
III	✓	✗	43.5	83.2	7.2
IV	✓	✓	**46.8**	**99.8**	**12.3**

Table 8. Oracle results on motion category and mask.

ID	Motion	Mask	$\mathcal{J}\&\mathcal{F}$	T-acc	N-acc
I	✓	✗	63.6	100.0	100.0
II	✗	✓	74.3	73.2	100.0

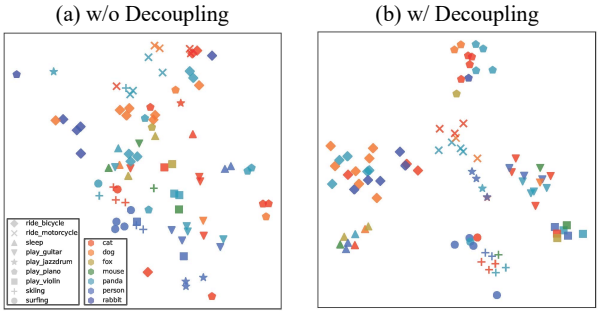
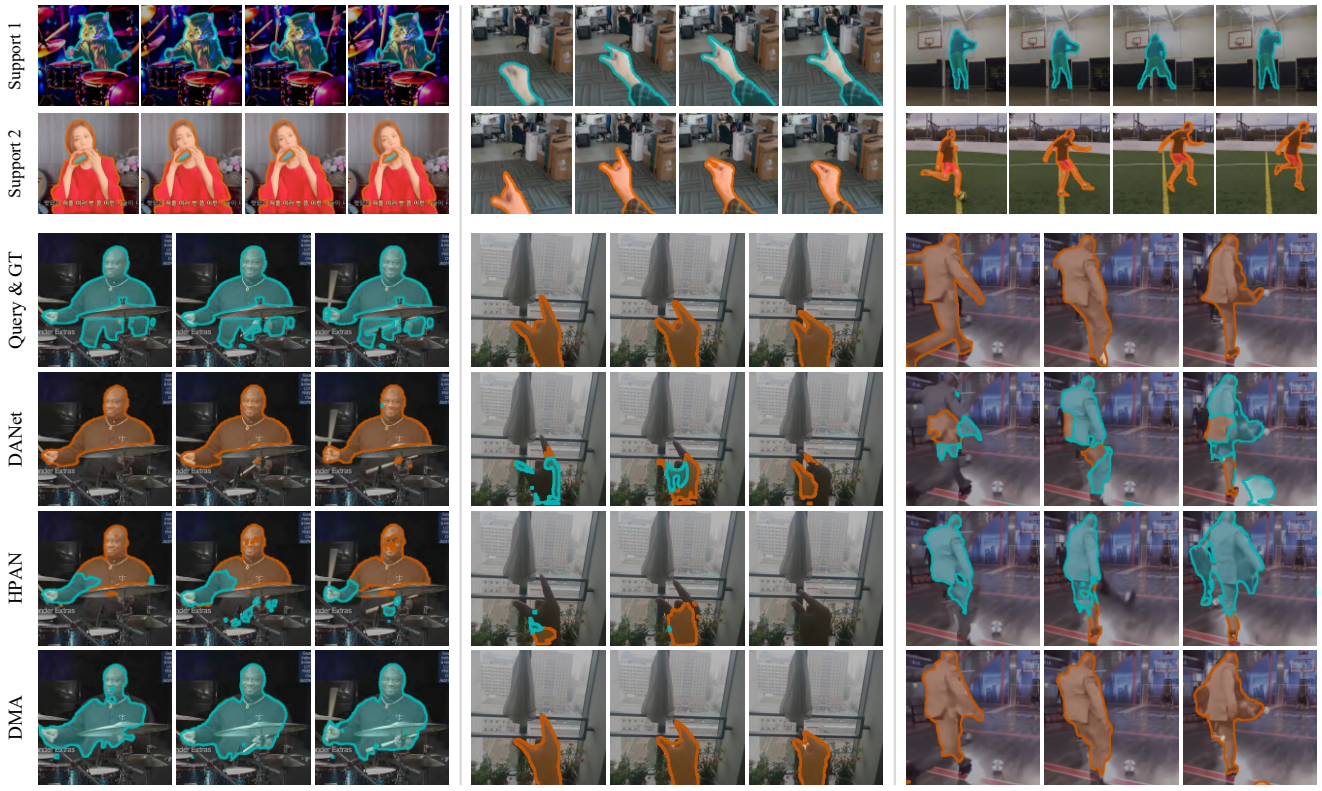


Figure 5. t-SNE [55] visualization of prototypes in our model (a) **w/o decoupling** and (b) **w/ decoupling**. Different colors and different shapes represent the object categories (*e.g.*, cat) and motion categories (*e.g.*, surfing), respectively. The proposed DMA effectively extracts the motion-centric prototypes and makes those having the same motions closer in feature space.

t-SNE Visualzation of Prototypes. As shown in Figure 5, we visualize the decoupled motion-appearance prototypes P_{dma} using t-SNE [55]. Without our DMA approach for prototype extraction, prototypes cluster according to object categories, *i.e.*, colors. In contrast, with our proposed DMA approach, prototypes cluster based on motion categories, *i.e.*, shapes, highlighting the effectiveness of our method in capturing motion-centric representations rather than appearance-based features.

5.4. Qualitative Results

Figure 6 presents several representative examples comparing our DMA with the baseline methods DANet [3] and HPAN [54]. In case (a), we showcase a challenging scenario where objects of different categories perform the same action: a cat playing the drums and a person playing the flute in the support videos while a person playing the drums in the query video. Baseline methods fail by segmenting based on the same object category of "person", whereas our method correctly segments the target based on the shared motion pattern, "playing the drums". This demonstrates



(a) Different categories with the same action (b) Supports with strong temporal correlation (c) Background scene misleading

Figure 6. Qualitative comparison of representative cases from MOVE between baseline methods, DANet [3] and HPAN [54], and our proposed DMA. (a) shows different object categories of "*cat*" (Support 1) and "*person*" (Query) performing the same action, "*playing drums*". (b) presents temporally correlated motions: fingers transitioning "*from pinching to opening*" (Support 1) and "*from opening to pinching*" (Support 2 & Query videos). (c) is a misleading background in the Query video, playing "*football*" on the "*basketball court*".

the effectiveness of our DMA design in prioritizing motion cues over object class identity. In case (b), we highlight a scenario with strong temporal correlations between frames in the support set: fingers transitioning from pinching to opening and from opening to pinching. While baseline methods struggle with fine-grained action discrimination due to insufficient temporal modeling, our proposed method effectively captures subtle temporal dependencies, leading to precise motion recognition and object segmentation. In case (c), our model correctly segments the target object, whereas the baseline methods are misled by the background context of playing "football" on the "basketball court", and fail to capture the specific motion category. These examples demonstrate the superiority of our approach in few-shot motion understanding. Additional failure cases are provided in the supplementary material.

6. Conclusion

We introduce MOVE, a new benchmark for motion-guided few-shot video object segmentation. Unlike existing FSVOS datasets that segment objects based on object categories, the proposed MOVE emphasizes temporal dynamics by segmenting objects according to motion categories that correlate with support and query videos. Experimental results show that MOVE poses significant challenges to current state-of-the-art methods, motivating us to propose DMA that decouples motion and appearance prototypes for more robust and effective motion prototype extraction. MOVE provides a foundation for advancing research in motion-centric few-shot video understanding and temporal modeling for segmentation tasks.

Future Directions. The MOVE benchmark opens up several promising research directions that need further investigation. We highlight some key areas for future exploration: i) decomposing complex motions into meta-motions for more general and efficient motion prototype learning, ii) modeling relational motions that involve interactions between multiple objects, iii) improving fine-grained motion discrimination through extracting more robust motion prototypes, iv) handling long-term temporal motions spanning multiple seconds through efficient temporal modeling, and v) learning discriminative background prototypes to suppress false positives in complex scenes better.

Acknowledgement. This project was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62472104.

References

[1] Goutam Bhat, Felix Järemo Lawin, Martin Danelljan, Andreas Robinson, Michael Felsberg, Luc Van Gool, and Radu Timofte. Learning what to learn for video object segmentation. In *Eur. Conf. Comput. Vis.*, 2020. 2

[2] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[3] Haoxin Chen, Hanjie Wu, Nanxuan Zhao, Sucheng Ren, and Shengfeng He. Delving Deep Into Many-to-Many Attention for Few-Shot Video Object Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 2, 3, 5, 6, 7, 8

[4] Jun Chen, Ming Hu, Darren J Coker, Michael L Berumen, Blair Costelloe, Sara Beery, Anna Rohrbach, and Mohamed Elhoseiny. Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 3

[5] Lei Chen, Zhan Tong, Yibing Song, Gangshan Wu, and Limin Wang. Efficient video action detection with token dropout and context refinement. In *Int. Conf. Comput. Vis.*, 2023. 3

[6] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *Eur. Conf. Comput. Vis.*, 2022. 2

[7] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Adv. Neural Inform. Process. Syst.*, 2021. 2

[8] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 2

[9] Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. Haa500: Human-centric atomic action dataset with curated videos. In *Int. Conf. Comput. Vis.*, 2021. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009. 5

[11] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *Int. Conf. Comput. Vis.*, 2023. 2, 3, 5, 6

[12] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *Int. Conf. Comput. Vis.*, 2023. 2, 5

[13] Henghui Ding, Song Tang, Shuting He, Chang Liu, Zuxuan Wu, and Yu-Gang Jiang. Multimodal referring segmentation: A survey. *arXiv*, 2025. 2

[14] Henghui Ding, Kaining Ying, Chang Liu, Shuting He, Yu-Gang Jiang, Philip HS Torr, and Song Bai. MOSEv2: A more challenging dataset for video object segmentation in complex scenes. *arXiv*, 2025. 2

[15] Hongdong Li Dongxu Li. Wlasl (world level american sign language) video, 2022. 3

[16] Fan, Qi, Tang, Chi-Keung, Tai, and Yu-Wing. Few-Shot Video Object Detection. In *Eur. Conf. Comput. Vis.*, 2022. 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 4, 5, 6

[18] Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 2, 3

[19] Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbab, Chaoyi Pan, et al. Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. *arXiv preprint arXiv:2502.01143*, 2025. 1

[20] Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. In *Int. Conf. Comput. Vis.*, 2023. 2

[21] Lingyi Hong, Zhongying Liu, Wenchao Chen, Chenzhi Tan, Yuang Feng, Xinyu Zhou, Pinxue Guo, Jinglun Li, Zhaoyu Chen, Shuyong Gao, et al. Lvos: A benchmark for large-scale long-term video object segmentation. *arXiv preprint arXiv:2404.19326*, 2024. 2

[22] Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025. 3

[23] HE Jian and WANG Weidong. Visual recognition of chinese traffic police gestures based on spatial context and temporal features. *Acta Electronica Sinica*, 2020. 3

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3, 5

[25] Seongchan Kim, Woojeong Jin, Sangbeom Lim, Heeji Yoon, Hyunwook Choi, and Seungryong Kim. Referring video object segmentation via language-aligned track selection. *arXiv preprint arXiv:2412.01136*, 2024. 3

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Int. Conf. Comput. Vis.*, 2023. 2

[27] Yu Kong, Yunde Jia, and Yun Fu. Learning human interaction by interactive phrases. In *Eur. Conf. Comput. Vis.*, 2012. 3

[28] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. *arXiv preprint arXiv:1911.06644*, 2019. 3

[29] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video

database for human motion recognition. In *Int. Conf. Comput. Vis.*, 2011. 3

[30] Ge Li, Hanqing Sun, Aiping Yang, Jiale Cao, and Yanwei Pang. Motion expressions guided video segmentation via effective motion information mining. *IEEE Trans. Emerg. Topics Comput. Intell.*, 2025. 2, 3

[31] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Int. Conf. Comput. Vis.*, 2021. 3

[32] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Int. Conf. Comput. Vis.*, 2021. 3

[33] Yongkang Li, Tianheng Cheng, Wenyu Liu, and Xinggang Wang. Mask-adapter: The devil is in the masks for open-vocabulary segmentation. *arXiv preprint arXiv:2412.04533*, 2024. 7

[34] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 4

[35] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 5

[36] Chang Liu, Xudong Jiang, and Henghui Ding. Primitivenet: decomposing the global constraints for referring segmentation. *Visual Intelligence*, 2024. 2

[37] Nian Liu, Kepan Nan, Wangbo Zhao, Yuanwei Liu, Xiwen Yao, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Junwei Han, and Fahad Shahbaz Khan. Multi-grained Temporal Prototype Learning for Few-shot Video Object Segmentation. In *Int. Conf. Comput. Vis.*, 2023. 1, 2, 3

[38] Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, et al. Convbench: A multi-turn conversation evaluation benchmark with hierarchical ablation capability for large vision-language models. In *Adv. Neural Inform. Process. Syst. D&B*, 2024. 3

[39] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 4, 5, 6

[40] Naisong Luo, Yuan Wang, Rui Sun, Guoxin Xiong, Tianzhu Zhang, and Feng Wu. Exploring the Better Correlation for Few-Shot Video Object Segmentation. *IEEE Trans. Circuits Syst. Video Technol.*, 2024. 1, 2, 3

[41] Naisong Luo, Yuan Wang, Rui Sun, Guoxin Xiong, Tianzhu Zhang, and Feng Wu. Holistic prototype attention network for few-shot video object segmentation. *IEEE Trans. Circuits Syst. Video Technol.*, 2024. 3

[42] Binjie Mao, Xiyan Liu, Linsu Shi, Jiazhong Yu, Fei Li, and Shiming Xiang. Few-shot video object segmentation with prototype evolution. *Neural Computing and Applications*, 2024. 1, 2

[43] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Int. Conf. Comput. Vis.*, 2019. 2

[44] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *Int. Conf. Comput. Vis.*, 2021. 3

[45] Chirag Parikh, Rohit Saluja, CV Jawahar, and Ravi Kiran Sarvadevabhatla. Idd-x: A multi-view dataset for ego-relative important object localization and explanation in dense and unstructured traffic. In *IEEE Int. Conf. Robot. Autom.*, 2024. 3

[46] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2

[47] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[48] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[49] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *ACM Int. Conf. Multimedia*, 2007. 3

[50] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 3

[51] Mennatullah Siam. Temporal transductive inference for few-shot video object segmentation. *Int. J. Comput. Vis.*, 2025. 1, 3, 6

[52] Mennatullah Siam, Konstantinos G Derpanis, and Richard P Wildes. Temporal transductive inference for few-shot video object segmentation. *arXiv preprint arXiv:2203.14308*, 2022. 1, 2, 3

[53] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3

[54] Yin Tang, Tao Chen, Xiruo Jiang, Yazhou Yao, Guo-Sen Xie, and Heng-Tao Shen. Holistic prototype attention network for few-shot video object segmentation. *IEEE Trans. Circuit Syst. Video Technol.*, 2023. 1, 2, 3, 5, 6, 7, 8

[55] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *J. Mach. Learn. Res.*, 2008. 7

[56] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Int. Conf. Comput. Vis.*, pages 3551–3558, 2013. 3

[57] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Int. Conf. Comput. Vis.*, 2019. 1

[58] Huaxin Xiao, Jiashi Feng, Guosheng Lin, Yu Liu, and Maojun Zhang. Monet: Deep motion exploitation for video

object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[59] Qianxiong Xu, Wenting Zhao, Guosheng Lin, and Cheng Long. Self-calibrated cross attention network for few-shot segmentation. In *Int. Conf. Comput. Vis.*, 2023. 1, 5, 6

[60] Qianxiong Xu, Guosheng Lin, Chen Change Loy, Cheng Long, Ziyue Li, and Rui Zhao. Eliminating feature ambiguity for few-shot segmentation. In *Eur. Conf. Comput. Vis.*, 2024. 1

[61] Shilin Yan, Renrui Zhang, Ziyu Guo, Wenchao Chen, Wei Zhang, Hongyang Li, Yu Qiao, Hao Dong, Zhongjiang He, and Peng Gao. Referred by multi-modality: A unified temporal transformer for video object segmentation. In *AAAI*, 2024. 2

[62] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K Katsaggelos. Efficient video object segmentation via network modulation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[63] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Int. Conf. Comput. Vis.*, 2019. 3, 5

[64] Le Yang, Houwen Peng, Dingwen Zhang, Jianlong Fu, and Junwei Han. Revisiting anchor mechanisms for temporal action localization. *IEEE Trans. Image Process.*, 2020. 3

[65] Kaining Ying, Zhenhua Wang, Cong Bai, and Pengfei Zhou. Isda: Position-aware instance segmentation with deformable attention. In *IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022. 5

[66] Kaining Ying, Qing Zhong, Weian Mao, Zhenhua Wang, Hao Chen, Lin Yuanbo Wu, Yifan Liu, Chengxiang Fan, Yunzhi Zhuge, and Chunhua Shen. CTVIS: Consistent training for online video instance segmentation. In *Int. Conf. Comput. Vis.*, 2023. 5

[67] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. MMT-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. In *Int. Conf. Mach. Learn.*, 2024. 3

[68] Kaining Ying, Henghui Ding, Guangquan Jie, and Yu-Gang Jiang. Towards omnimodal expressions and reasoning in referring audio-visual segmentation. In *Int. Conf. Comput. Vis.*, 2025. 2

[69] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei. Few-shot segmentation via cycle-consistent transformer. *Adv. Neural Inform. Process. Syst.*, 2021. 1, 6

[70] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. Egogesture: a new dataset and benchmark for egocentric hand gesture recognition. *IEEE Trans. Multimedia*, 2018. 3

[71] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *Int. Conf. Comput. Vis.*, 2021. 3

[72] Tianfei Zhou, Fatih Porikli, David J Crandall, Luc Van Gool, and Wenguan Wang. A survey on deep learning technique for video segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 1

[73] Wangjiang Zhu, Jie Hu, Gang Sun, Xudong Cao, and Yu Qiao. A key volume mining deep framework for action recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 3