REMOM: RELATION-AWARE MOTION MATCHING FOR MOTION-GUIDED FEW-SHOT VIDEO OBJECT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Motion-guided few-shot video object segmentation (M-FSVOS) aims to segment objects in a query video that exhibit motion patterns defined by a few support videos. A key challenge of M-FSVOS lies in support-defined motion correspondence, *i.e.*, reliably matching motion across support and query videos despite variations in appearance and execution style. Although state-of-the-art video object segmentation models such as SAM 3 excel at object discovery, segmentation, and tracking, their object representations rely heavily on appearance cues and thus fall short in the M-FSVOS task. To tackle this challenge, we introduce a novel model named Relation-aware Motion Matching (ReMoM), which formulates M-FSVOS as a motion-centric masklet matching problem. Under this formulation, ReMoM performs support-conditioned motion matching through explicit track-to-way relation modeling over high-quality masklets extracted by SAM 3. Additionally, it strengthens the temporal evidence of the learned motion relations through counterfactual temporal enhancement and incorporates interaction-relevant context from surrounding objects to resolve interaction-dependent ambiguities. Extensive experiments on the MOVE and A2D-Motion benchmarks demonstrate the effectiveness of ReMoM, which significantly outperforms state-of-the-art models by 17.5% on MOVE.

1 INTRODUCTION

Video object segmentation (VOS) aims to segment the objects of interest in a video, and serves as a fundamental task for a wide range of real-world applications, including autonomous driving, surveillance, robotics and augmented reality (Oh et al., 2019; Cheng & Schwing, 2022). However, existing VOS methods require large amounts of training data and can hardly recognize unseen classes. To address this limitation, few-shot video object segmentation (FSVOS) has been proposed (Chen et al., 2021; Siam, 2025), which aims to segment objects of unseen classes in videos using only a few annotated examples. By requiring minimal supervision, FSVOS lowers the demand for large-scale labeled data and enables fast adaptation to novel object classes. However, existing FSVOS models define target categories through object appearance in static support images, rather than through motion patterns demonstrated across videos.

To address this issue, motion-guided few-shot video object segmentation (M-FSVOS) (Ying et al., 2025) aims to segment the objects in a query video that exhibit motion patterns defined by a few support videos. Unlike FSVOS, M-FSVOS specifies the target with exemplar videos instead of static images, and the segmentation is guided by the motion patterns, regardless of object category. As illustrated in Fig. 1, a support video showing a car overtaking on the right defines the target motion as *overtaking*. The model is then required to segment the motorcycle performing the same motion in the query video, irrespective of its object category.

In principle, M-FSVOS presents a significant challenge: the model is required to recognize and match motion patterns across support and query videos despite variations in object appearance and execution style. As illustrated by the first example in Fig. 1, a car in the support video and a motorcycle in the query video may perform the same overtaking motion, illustrating the appearance variation. In the second example, the same push-up motion is performed with two hands in the support video and one

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

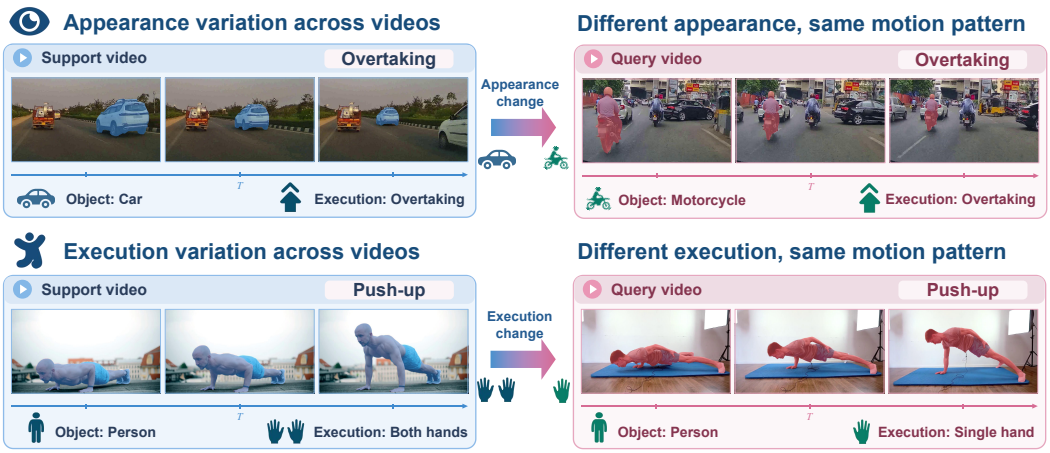

Figure 1: Motion-guided few-shot video object segmentation requires support-defined motion correspondence despite variations in appearance and execution. Support videos (left) and query videos (right) illustrate two motion classes: *overtaking* (top) and *push-up* (bottom). The same motion is matched across appearance variations (car *vs*. motorcycle) and execution variations (push-ups with both hands *vs*. a single hand), respectively.

hand in the query video, illustrating the execution variation. In both cases, models should rely on the underlying motion pattern rather than appearance or execution style.

Recently, promptable segmentation models (*e.g*., SAM 2, SAM 3) (Kirillov et al., 2023; Ravi et al., 2024; Carion et al., 2025) have emerged as powerful foundation models for video segmentation. These models excel at producing high-quality object tracks (masklets) across video frames together with their object-token sequences that implicitly encode object trajectories. Despite their remarkable performance, these models focus primarily on object-centric understanding and lack explicit motion understanding capabilities. Thus, they fall short in M-FSVOS where target categories are instead defined by motion patterns.

To tackle the above challenge, we propose ReMoM (Relation-aware Motion Matching), which formulates M-FSVOS as motion-centric masklet matching. To learn the motion relation between support and query videos, ReMoM leverages the object-token sequences produced by SAM 3, and comprises three complementary components: (i) *Track-to-Way Motion Matching*, which explicitly models the relation between query tracks and support motion exemplars, capturing fine-grained motion agreement and discrepancy for support-specific discrimination; (ii) *Counterfactual Temporal Enhancement*, which strengthens temporal evidence of relational representations by contrasting the factual view of the query with an appearance-preserving counterfactual view; (iii) *Interaction-Aware Contextualization*, which injects interaction-relevant object context to discriminate interaction-dependent motions by selectively aggregating contextual information from the other query tracks.

In our experiments, we evaluate ReMoM on the MOVE (Ying et al., 2025) benchmark, where it outperforms previous state-of-the-art models by 17.5% in terms of $\mathcal{J}\&\mathcal{F}$. To test generalization beyond MOVE, we further construct a new benchmark, named A2D-Motion, based on the A2D dataset (Xu et al., 2015). On this benchmark, ReMoM consistently outperforms strong baselines across folds. Ablation studies demonstrate that each of the three components contributes to performance. We additionally design motion-centric analyses which probe temporal-order sensitivity and appearance-motion shortcuts to verify that the model's decisions are driven by motion rather than appearance, and both qualitative and quantitative results show that ReMoM effectively captures the motion patterns. Further experiments and results are presented in the supplementary material.

In summary, our contributions are listed as follows:

- Unlike previous works, we reformulate M-FSVOS as a motion-centric masklet matching problem, providing a natural framework for equipping SAM 3 with motion understanding capabilities tailored to M-FSVOS.
- We propose ReMoM, which explicitly models the relations between query tracks and support motion exemplars for support-specific motion discrimination. Additionally, ReMoM strengthens the temporal evidence through counterfactual temporal enhancement and incorporates interaction-relevant context to resolve interaction-dependent ambiguities.

- We construct A2D-Motion, a new action-generalization benchmark that recasts A2D in the M-FSVOS setting. Extensive experiments on MOVE and A2D-Motion demonstrate the effectiveness of the proposed ReMoM.

2 Related Work

Video object segmentation (VOS) (Perazzi et al., 2016; Ding et al., 2025; Hong et al., 2024; Xiao et al., 2018; Cheng et al., 2024; Caelles et al., 2017; Bhat et al., 2020) aims to segment the target objects in a video, given the object mask in the initial frame. Recent methods typically use memory-based designs that store and recall historical frame features (Oh et al., 2019; Cheng & Schwing, 2022; Cheng et al., 2021). Referring video object segmentation (RVOS) instead specifies the target with a natural-language expression (Seo et al., 2020; Khoreva et al., 2018; Ding et al., 2023; He & Ding, 2024). Recent query-based models segment the referred object with object query tokens (Wu et al., 2022; Botach et al., 2022; Ding et al., 2023). However, these methods face limitations when dealing with novel or complex motions that are difficult to describe in language.

Few-shot video object segmentation (FSVOS) (Chen et al., 2021; Siam, 2025; Liu et al., 2023; Luo et al., 2024) aims to segment objects of categories unseen during training in query videos, given only a few annotated support images and masks. Related few-shot learning and image segmentation approaches use support-query matching (Vinyals et al., 2016; Min et al., 2021), prototypes (Snell et al., 2017; Wang et al., 2019), and prior-guided features (Tian et al., 2020; Wang et al., 2024). Within FSVOS, many-to-many attention links support images to query frames (Chen et al., 2021), while temporal transductive inference regularizes predictions across query frames (Siam, 2025). Multi-grained temporal prototypes further capture local and long-term video context (Liu et al., 2023). Although FSVOS methods exploit temporal consistency in query videos, their target categories remain defined by object appearance rather than support-demonstrated motion, motivating a motion-centric formulation.

Motion-guided few-shot video object segmentation (M-FSVOS) aims to segment objects in a query video that exhibit motion patterns defined by a few support videos. Recently, Ying et al. (2025) proposed the MOVE benchmark and the Decoupled Motion-Appearance network (DMA), which decouples appearance and motion into prototypes that guide mask decoding and are matched by cosine similarity. Our proposed ReMoM significantly differs from DMA in two aspects. First, we reformulate M-FSVOS as track matching: whereas DMA uses the decoupled prototypes to guide a mask decoder, we directly match each query track against each support motion and segment by selecting the matched tracks. Second, we replace the cosine similarity with a learned pairwise relation encoder that explicitly captures the fine-grained agreement and discrepancy between query tracks and support motion exemplars. We also supervise it with a contrastive objective that shuffles temporal order while preserving appearance, so that the model will not rely on appearance shortcuts.

Segment anything models. The SAM family (Kirillov et al., 2023; Ravi et al., 2024; Carion et al., 2025) provides foundation models for promptable image and video segmentation. Due to their strong segmentation ability, these models have recently been adapted to few-shot segmentation: some encode support features as memory and match query features against it (Xu et al., 2025), while others combine vision-language models with LLM-driven semantic cues (Karimi & Poullis, 2025). These works, however, are trained to segment a target object by its identity or appearance. Consequently, these models cannot determine which query object exhibits the motion demonstrated by the support videos, which is exactly what M-FSVOS requires.

3 Method

3.1 Task Formulation

Motion-guided few-shot video object segmentation (M-FSVOS) (Ying et al., 2025) aims to segment objects by the motion patterns they exhibit. Specifically, given a query video and a few support videos demonstrating distinct motion patterns, the model must predict the mask of the query object that performs the same motion as one of the support videos, regardless of its category.

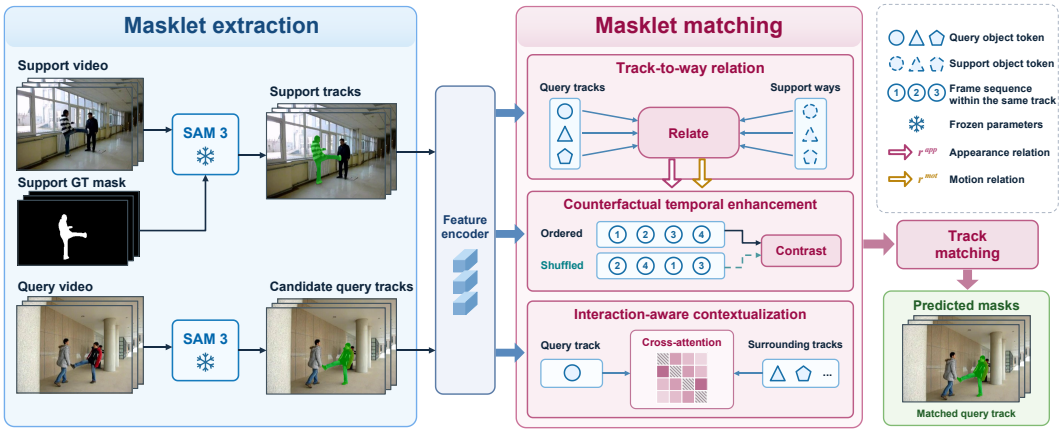

Figure 2: An overview of ReMoM. Frozen SAM 3 first extracts query tracks. Encoded features undergo track-to-way relation modeling, counterfactual temporal enhancement, and interaction-aware contextualization. Candidate query tracks are matched to support ways to produce the segmentation.

We follow the W-way-K-shot episodic setting, where each episode consists of a support set $\mathcal{S}$ and a query set $\mathcal{Q}$. The support set $\mathcal{S} = \{(V^s_{w,k}, M^s_{w,k})\}^{W,K}_{w=1,k=1}$ has W distinct motion patterns (categories), each demonstrated by K support videos, where $V^s_{w,k}$ is the k-th support video of way w and $M^s_{w,k}$ is the mask sequence of the object performing that way's motion. The query set $\mathcal{Q} = \{I^q_t\}^T_{t=1}$ is a video sequence of T frames. Given $\mathcal{S}$ and $\mathcal{Q}$, the goal is to predict the query mask sequence $\hat{M}^q = \{\hat{M}^q_t\}^T_{t=1}$ that corresponds to one of the W motion ways, *i.e.*, the mask of the object performing the same motion as a support way. The dataset is split into a base training set and a novel test set with disjoint motion categories. The model is trained on episodes sampled from the training set and evaluated on episodes from the test set.

3.2 Motion-centric Masklet Matching

To address M-FSVOS, we reformulate the task as motion-centric masklet matching over the object tokens extracted by SAM 3 (Carion et al., 2025). Specifically, we first use SAM 3 to produce a set of candidate masklets in the query video, and then perform support-conditioned motion matching. In support-conditioned motion matching, each query masklet is matched to a support way based on its motion pattern (or to a null way), and the final query mask is assembled from the matched masklets. This reformulation offers two advantages: 1) By decomposing M-FSVOS into masklet extraction and masklet matching, we can directly leverage the strong pixel-level object discovery and tracking capabilities of SAM 3. 2) More importantly, our masklet matching formulation provides a natural way to equip SAM 3 with motion understanding capabilities tailored specifically for M-FSVOS.

As illustrated in Fig. 2, ReMoM first extracts motion and appearance features from the object tokens with a feature encoder. Based on these features, we contribute three novel modules to conduct motion-centric masklet matching: namely (1) a track-to-way relation encoder for extracting the motion relation between tracks and ways, (2) a counterfactual temporal enhancement module for modeling the temporal dynamics by contrasting ordered and shuffled sequences, and (3) an interaction-aware contextualization module for attending to the surrounding query tracks. Finally, a prediction head matches each query masklet to a support way or the null way. We detail the modules in Section 3.3.

SAM 3-based Track Generation. We use a frozen SAM 3 (Carion et al., 2025) to extract masklets. Specifically, given a video and a prompt, SAM 3 tracks the prompted object and outputs a masklet, *i.e.*, the object's mask propagated across all frames. We also refer to each masklet as an object track. For each masklet, SAM 3 also outputs an object token sequence, whose per-frame object pointer $O_{i,t} \in \mathbb{R}^D$ encodes the object's identity and trajectory. We concatenate these pointers into an object-token sequence $O_i = [O_{i,1}; \ldots; O_{i,T}] \in \mathbb{R}^{T\times D}$, which encapsulates the object's appearance and trajectory over time. To produce candidate masklets, we prompt the query video with a grid of points at frames sampled at a fixed interval, obtaining N masklets and their corresponding object tokens $O^q = \{O^q_i\}^N_{i=1}$. Each support video $V^s_{w,k}$ is instead prompted by its ground-truth mask, yielding one anchored support object-token sequence $O^s_{w,k} \in \mathbb{R}^{T_{w,k}\times D}$. For supervision, we assign each query

masklet a pseudo label by its IoU with the ground-truth mask of the target motion: masklets with an IoU above a threshold are positive (the set $\mathcal{P}$) and the rest are negative (the set $\mathcal{N}$).

3.3 Relation-aware Motion Matching

Given the N query masklets and the $W \times K$ support masklets with their object token sequences O^q and O^s, we perform M-FSVOS by matching each query track against every support way. Our proposed ReMoM takes the query object tokens O^q and the support object tokens O^s as input, and outputs a score vector over the W support ways plus a null way. The score vector for the i-th query track is denoted by s_i, which is given as follows:

$$s_i = \mathrm{RM}\big(O_i^q;\ \{O_{w,k}^s\}_{w=1,k=1}^{W,K}\big) \in \mathbb{R}^{W+1}, \tag{1}$$

where $\mathrm{RM}(\cdot)$ denotes the relation-aware motion matching model.

Feature encoder. To extract motion features from the query and support tracks, we design a shared motion encoder Φ_{mot}. As motion is reflected in how the tokens change over time rather than in their absolute values, the motion encoder takes temporal differences between adjacent frames as input. The motion encoder is formulated as follows:

$$D_{i,t}^q = L_2(O_{i,t+1}^q) - L_2(O_{i,t}^q), \qquad \boldsymbol{m}_i^q = \Phi_{\mathrm{mot}}(O_i^q) = \mathrm{TCN}\big(D_i^q\big), \tag{2}$$

where $L_2(\cdot)$ denotes ℓ_2 normalization, and $\mathrm{TCN}(\cdot)$ denotes a stack of residual dilated temporal convolutions (Bai et al., 2018). The support features are aggregated into way features $\boldsymbol{m}_w^s$. Appearance also helps motion discrimination, as it supplies the spatial layout and the moving parts of the object. We therefore additionally introduce a shared appearance encoder Φ_{app}. It performs self-attention over the object tokens with a temporal transformer (Vaswani et al., 2017) to encode an appearance feature $\boldsymbol{a}_i^q = \Phi_{\mathrm{app}}(O_i^q) \in \mathbb{R}^D$, with the support features aggregated into $\boldsymbol{a}_w^s$. To ensure the extraction of motion information, we introduce a W-way auxiliary loss over the positive masklets for supervision, $\mathcal{L}_{\mathrm{mw}} = \mathrm{CE}(\mathrm{aux}(\boldsymbol{m}_i^q; \{\boldsymbol{m}_w^s\}_{w=1}^W),\ y_i)$, where $\mathrm{aux}(\cdot)$ is an auxiliary classifier that outputs the probability distribution of the query track i over the W support ways.

Track-to-way relation encoder. To learn the correspondence between query tracks and support ways, we propose to model the track-to-way relation to capture fine-grained motion agreement and discrepancy for support-specific discrimination. This design takes inspiration from the support-defined motion nature of M-FSVOS, where motion categories are specified by support motion exemplars. Accordingly, recognizing a query motion requires explicitly comparing it against the demonstrated support motions. Specifically, we instantiate the track-to-way relation with a motion relation module g_θ^{mot}, based on the motion features extracted by the motion encoder. It takes as input the concatenation of the query motion feature, the way motion feature, and their absolute difference, where the absolute-difference term captures fine-grained agreement and discrepancy: a small difference indicates agreement, while a large one pinpoints where the track and the way diverge. Since the query may contain no object performing any of the support motions, we also append a learnable null-way embedding $\boldsymbol{m}_{\mathrm{null}}$ to the way features. The module is formulated as follows:

$$\boldsymbol{r}_{ij}^{\mathrm{mot}} = g_\theta^{\mathrm{mot}}\big([\boldsymbol{m}_i^q, \boldsymbol{m}_j^s, |\boldsymbol{m}_i^q - \boldsymbol{m}_j^s|]\big), \tag{3}$$

where i indexes the query track and j ranges over the W support ways and the null way. Additionally, we propose an appearance relation encoder g_θ^{app} as an auxiliary, defined analogously on the appearance features: $\boldsymbol{r}_{ij}^{\mathrm{app}} = g_\theta^{\mathrm{app}}([\boldsymbol{a}_i^q, \boldsymbol{a}_j^s, |\boldsymbol{a}_i^q - \boldsymbol{a}_j^s|])$.

Counterfactual temporal enhancement. In principle, motion is tied to temporal order, as changing the order of a motion sequence may alter its semantics. For example, reversing the temporal order of a leftward motion turns it into a rightward motion. To strengthen temporal evidence of the motion relation by enhancing the temporal sensitivity, we propose a counterfactual temporal enhancement module, which enhances temporal information by contrasting the factual view of the query with an appearance-preserving counterfactual view. The module is formulated as follows:

$$\begin{aligned} \boldsymbol{m}_i^{\mathrm{cf}} &= \Phi_{\mathrm{mot}}\big(\mathcal{T}(O_i^q)\big), & \boldsymbol{\delta}_{ij} &= \boldsymbol{r}_{ij}^{\mathrm{mot}} - \boldsymbol{r}_{ij}^{\mathrm{mot,cf}}, \\ \boldsymbol{h}_{ij} &= \phi\big(\boldsymbol{\delta}_{ij},\ \boldsymbol{r}_{ij}^{\mathrm{app}}\big), & \tilde{\boldsymbol{r}}_{ij} &= \boldsymbol{r}_{ij}^{\mathrm{mot}} + \alpha\, \boldsymbol{h}_{ij}, \end{aligned} \tag{4}$$

where $\mathcal{T}(\cdot)$ denotes the temporal shuffle and $\boldsymbol{r}_{ij}^{\mathrm{mot,cf}}$ is the counterfactual motion relation obtained by matching the counterfactual motion feature $\boldsymbol{m}_i^{\mathrm{cf}}$ against the support way features. Since shuffling preserves the appearance but disrupts the temporal order of object tokens, we construct the counterfactual motion difference, $\boldsymbol{\delta}_{ij}$, which isolates exactly what the temporal order contributes. We obtain the residual $\boldsymbol{h}_{ij}$ with a gated residual encoder ϕ, where the appearance relation $\boldsymbol{r}_{ij}^{\mathrm{app}}$ serves as a gate to condition the enhancement on the moving parts. Finally, we enhance the motion relation by the addition of the residual, where α is a non-negative learnable scale.

Interaction-aware contextualization. In M-FSVOS, many interaction-dependent motions cannot be fully characterized by the target object's own trajectory, as their semantics also depend on the objects involved in the interaction. For example, similar target trajectories may correspond to different motions depending on whether the target interacts with a bike or a horse. In such cases, the target track alone provides insufficient evidence, making surrounding-object context important for motion discrimination. Inspired by this observation, we propose an interaction-aware contextualization module, which injects interaction-relevant object context to improve the discrimination of interaction-dependent motions. For every candidate way j, it builds a query from the appearance and motion features of the current pair and cross-attends to the appearance features of the other query tracks, yielding a context feature,

$$\boldsymbol{c}_{ij} = \mathrm{Context}\Big(\big[\boldsymbol{a}_i^q, \boldsymbol{a}_j^s, \boldsymbol{m}_i^q, \boldsymbol{m}_j^s\big],\ \{\boldsymbol{a}_k^q\}_{k\neq i}\Big), \tag{5}$$

which injects the interaction evidence that the target's own track does not contain. We attend only to the appearance of the surrounding query tracks rather than to their motion, because the interaction evidence lies more in the identity of these objects. The query, in turn, combines the motion being evaluated with the appearance of the current object, so the context is conditioned on both the candidate query track and the support way. This contextualization module thereby helps discriminate interaction-dependent motions.

Finally, an MLP prediction head $\mathcal{F}$ maps the concatenation of the appearance relation, the enhanced motion relation, and the context feature to a scalar logit, $s_{ij} = \mathcal{F}([\boldsymbol{r}_{ij}^{\mathrm{app}}, \tilde{\boldsymbol{r}}_{ij}, \boldsymbol{c}_{ij}])$, which, after a softmax over the $W{+}1$ ways, gives the probability that track i matches way j or the null way.

3.4 Training Objective

Relation-aware matching loss. The main supervision is a cross-entropy over the $W{+}1$ ways, *i.e.*, the W support ways and the null way: positive masklets are pushed toward their way label y_i, and negative masklets toward the null way,

$$\mathcal{L}_{\mathrm{rm}} = \frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} \mathrm{CE}(s_i, y_i) + \lambda_{\mathrm{null}}\frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}} \mathrm{CE}(s_i, \mathrm{null}), \tag{6}$$

where $s_i \in \mathbb{R}^{W+1}$ denotes the logit vector for the i-th masklet.

Motion-aware contrastive loss. To supervise the counterfactual temporal enhancement of the motion relation, we introduce a motion-aware contrastive loss to force the model to be more confident about the factual view than about the counterfactual view. For the counterfactual view, we only replace the motion relation with the counterfactual one to formulate the counterfactual score,

$$s_{ij}^{\mathrm{cf}} = \mathcal{F}\big([\boldsymbol{r}_{ij}^{\mathrm{app}},\ \boldsymbol{r}_{ij}^{\mathrm{mot,cf}},\ \boldsymbol{c}_{ij}]\big), \tag{7}$$

where $\mathcal{F}$ is the same prediction head as in the factual view. For a positive masklet i, we measure its confidence by the decision margin, namely the difference between the target logit and the log-sum-exp over the competing ways. Mathematically, the decision margins are formulated as follows:

$$d_i^{\mathrm{f}} = s_{i,y_i} - \log\sum\nolimits_{j\neq y_i} e^{s_{ij}}, \qquad d_i^{\mathrm{cf}} = s_{i,y_i}^{\mathrm{cf}} - \log\sum\nolimits_{j\neq y_i} e^{s_{ij}^{\mathrm{cf}}}. \tag{8}$$

A large margin indicates that the target way y_i clearly surpasses all alternatives. Since shuffling destroys temporal order, a model that exploits motion should satisfy $d_i^{\mathrm{f}} > d_i^{\mathrm{cf}}$. The loss pulls the gap $\Delta_i = d_i^{\mathrm{f}} - d_i^{\mathrm{cf}}$ above a margin m_0 with a softplus,

$$\mathcal{L}_{\mathrm{cf}} = \frac{\tau}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} \mathrm{softplus}\Big(\frac{m_0 - \Delta_i}{\tau}\Big), \tag{9}$$

Table 1: Main results on MOVE under the Overlapping Split (OS). We report the $\mathcal{J}\&\mathcal{F}$ metric in the 2-way-1-shot and 5-way-1-shot settings, averaged over the four folds.

		$\mathcal{J}\&\mathcal{F}$(2-way-1-shot)					$\mathcal{J}\&\mathcal{F}$(5-way-1-shot)				
Method	Backbone	Mean	Fold 1	Fold 2	Fold 3	Fold 4	Mean	Fold 1	Fold 2	Fold 3	Fold 4
HPAN (Tang et al., 2024)	ResNet-50	44.4	48.4	45.2	43.4	40.8	34.0	37.6	34.8	34.6	29.1
TTI (Siam, 2025)	ResNet-50	45.2	45.8	43.9	43.7	47.4	35.6	33.8	35.9	34.8	37.8
DANet (Chen et al., 2021)	ResNet-50	45.4	41.4	44.7	47.1	48.2	25.4	27.4	23.5	25.8	25.0
DMA (Ying et al., 2025)	ResNet-50	50.1	51.2	46.2	54.3	48.6	40.2	40.7	38.9	41.3	39.7
DMA (Ying et al., 2025)	VSwin-T	51.5	51.1	48.6	56.3	50.0	41.4	41.5	39.8	42.7	41.1
SAM 3 Raw	SAM 3	30.6	30.3	30.6	30.5	31.0	30.9	31.4	30.7	29.9	31.6
SAM 3 + TimeAttn	SAM 3	60.0	59.8	58.8	62.9	58.7	49.1	49.2	49.4	51.2	46.6
SAM 3 + SOLA	SAM 3	59.1	56.5	58.1	61.8	59.8	45.9	44.5	46.9	47.0	45.2
ReMoM (Ours)	SAM 3	**69.0**	**68.2**	**67.5**	**70.2**	**70.0**	**58.7**	**57.8**	**58.0**	**59.3**	**59.6**

where $\mathrm{softplus}(x) = \log(1 + e^x)$ and τ is a temperature.

In summary, the overall training loss is formulated as

$$\mathcal{L} = \lambda_{\mathrm{rm}}\,\mathcal{L}_{\mathrm{rm}} + \lambda_{\mathrm{mw}}\,\mathcal{L}_{\mathrm{mw}} + \lambda_{\mathrm{cf}}\,\mathcal{L}_{\mathrm{cf}}, \tag{10}$$

where λ_{rm}, λ_{mw}, and λ_{cf} are weight coefficients.

4 Experiments

4.1 Datasets and evaluation metrics

We evaluate ReMoM on the MOVE benchmark (Ying et al., 2025) and on an action-generalization benchmark built on A2D (Xu et al., 2015), which we call A2D-Motion. Following MOVE (Ying et al., 2025), we report the $\mathcal{J}\&\mathcal{F}$ metric, the average of region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$, which jointly measure the segmentation quality in terms of region overlap and boundary accuracy.

MOVE. MOVE (Ying et al., 2025) contains 224 motion categories and evaluates whether a model segments objects by the motion they perform. For cross-validation, the categories are split into 4 folds with two strategies, the Overlapping Split (OS) and the Non-overlapping Split (NS). We follow the official protocol and report results on the *2-way-1-shot* and *5-way-1-shot* settings over the four folds under both OS and NS.

A2D-Motion. To further test generalization to motions unseen during training, we build an action-generalization benchmark on A2D (Xu et al., 2015). The eight regular actions are partitioned into four folds, each holding out two actions for testing and using the remaining six for training (Table 6). The "none" actions of A2D are evenly distributed across the four folds to construct negative episodes. Further construction details are provided in Appendix A.3.

4.2 Implementation details

The loss weights are λ_{rm}=1.0, λ_{null}=0.5, λ_{mw}=1.5, and λ_{cf}=0.4, with a margin m_0=0.2 and a temperature τ=0.1. We optimize with AdamW using per-component learning rates (3e$-$5 for the motion encoder, 1e$-$5 for the appearance encoder, and 1e$-$4 for the relation and prediction heads) and weight decay 1e$-$2. More implementation details are provided in Appendix A.1.

4.3 Main results

Baselines. We compare ReMoM with two groups of methods. The first group consists of three conventional few-shot video object segmentation (FSVOS) methods, HPAN (Tang et al., 2024), TTI (Siam, 2025), DANet (Chen et al., 2021), and a motion-guided few-shot video object segmentation (M-FSVOS) method, DMA (Ying et al., 2025). The second group consists of three SAM-based baselines, which operate on SAM 3 (Carion et al., 2025) object tokens. More detailed descriptions of the baselines are provided in Appendix A.2.

MOVE. As shown in Table 1, ReMoM achieves the best $\mathcal{J}\&\mathcal{F}$ on the Overlapping Split (OS), attaining 69.0% and 58.7% for the 2-way-1-shot and 5-way-1-shot settings, surpassing the strongest

Table 2: Main results on MOVE under the Non-overlapping Split (NS).

Method	Backbone	$\mathcal{J}\&\mathcal{F}$(2-way-1-shot)					$\mathcal{J}\&\mathcal{F}$(5-way-1-shot)				
		Mean	Fold 1	Fold 2	Fold 3	Fold 4	Mean	Fold 1	Fold 2	Fold 3	Fold 4
HPAN (Tang et al., 2024)	ResNet-50	39.1	49.1	34.9	40.0	32.3	30.2	35.3	28.5	30.3	26.6
TTI (Siam, 2025)	ResNet-50	43.6	47.2	33.4	50.0	43.9	32.7	35.0	29.8	35.7	30.2
DANet (Chen et al., 2021)	ResNet-50	44.6	48.4	36.9	49.5	43.6	29.9	29.5	24.0	31.0	35.3
DMA (Ying et al., 2025)	ResNet-50	46.0	47.8	37.9	48.0	50.3	34.7	35.6	31.5	37.0	34.6
DMA (Ying et al., 2025)	VSwin-T	49.0	54.4	37.4	48.5	55.9	35.4	37.9	29.9	36.6	37.2
SAM 3 Raw	SAM 3	30.2	39.7	25.4	26.0	29.8	28.4	33.4	25.9	27.6	26.6
SAM 3 + TimeAttn	SAM 3	55.4	51.2	50.3	57.0	63.3	40.3	38.6	39.7	43.6	39.4
SAM 3 + SOLA	SAM 3	53.2	54.6	49.9	52.8	55.7	37.5	39.5	37.9	39.2	33.5
ReMoM (Ours)	SAM 3	**62.1**	**56.0**	**56.5**	**67.5**	**68.4**	**47.3**	**41.7**	**45.2**	**54.5**	**47.6**

prior M-FSVOS method DMA by $+17.5\%$ and $+17.3\%$. ReMoM also ranks first on all four folds, demonstrating the effectiveness of our method. We further evaluate on the Non-overlapping Split (NS), a more challenging setting whose test categories come from unseen parent classes. As shown in Table 2, ReMoM still attains the best $\mathcal{J}\&\mathcal{F}$ (62.1% and 47.3%), surpassing DMA by $+13.1\%$ and $+11.9\%$, showing the robustness of our method. The trained SAM-based baselines consistently outperform prior FSVOS and M-FSVOS methods (*e.g.*, SAM 3 + TimeAttn reaches 60.0% *vs*. DMA's 51.5% in 2-way-1-shot OS), which validates the effectiveness of decoupling M-FSVOS into masklet extraction and masklet matching. ReMoM further surpasses all of them by a clear margin (*e.g.*, $+9.0\%$ over SAM 3 + TimeAttn in 2-way-1-shot OS), demonstrating the contribution of our three motion-matching modules.

A2D-Motion. Table 3 reports the results on A2D-Motion, where the two held-out actions are never seen during training. Since the A2D videos are sparsely annotated and appearance variations across videos are significant (*e.g.*, a bird flying *vs*. a ball flying, a car jumping *vs*. an adult jumping), the motion matching task is more challenging than on MOVE. Nevertheless, ReMoM consistently outperforms all baselines, with an average $\mathcal{J}\&\mathcal{F}$ of 45.3%, surpassing the strongest SAM-based baselines SAM 3 + SOLA (40.0%) and SAM 3 + TimeAttn (39.8%) as well as the best FSVOS method DANet (30.6%). The gain of $+5.3\%$ over SAM 3 + SOLA demonstrates the strong motion generalization ability of our ReMoM.

Table 3: A2D-Motion results.

Method	$\mathcal{J}\&\mathcal{F}$
HPAN (Tang et al., 2024)	20.8
TTI (Siam, 2025)	16.9
DANet (Chen et al., 2021)	30.6
DMA (Ying et al., 2025)	27.2
SAM 3 Raw	18.6
SAM 3 + TimeAttn	39.8
SAM 3 + SOLA	40.0
ReMoM (Ours)	**45.3**

4.4 Ablation Studies

Ablation on modules. As shown in Table 4, we start from a baseline that encodes the SAM 3 object tokens with a feature encoder and predicts the logits with a simple fusion head (65.5%). Adding the track-to-way relation encoder lifts $\mathcal{J}\&\mathcal{F}$ to 66.3% ($+0.8\%$), verifying the importance of modeling the track-to-way relation for support-conditioned matching. In addition, introducing the interaction-aware contextualization module further improves it to 67.9% ($+1.6\%$), showing that attending to surrounding objects helps disambiguate interaction-dependent motions. Furthermore, the counterfactual temporal enhancement reaches 69.0% ($+1.1\%$), indicating that enhancing the sensitivity to temporal order contributes to finer motion matching. The table also reports N-Acc, the accuracy of rejecting masklets that match no support motion, which is an auxiliary diagnostic beyond the primary metric $\mathcal{J}\&\mathcal{F}$. N-Acc is driven almost entirely by the relation and context features (11.1% to 23.2%), since these already determine whether a masklet matches any support motion. The counterfactual enhancement improves $\mathcal{J}\&\mathcal{F}$ while leaving the null-way decision comparable.

Ablation on auxiliary losses. As shown in Table 5, we ablate the two auxiliary losses using the full architecture, with all three modules included. Adding the motion-aware contrastive loss improves $\mathcal{J}\&\mathcal{F}$ from 67.4% to 67.9% ($+0.5\%$), showing that contrasting the factual and counterfactual views encourages the model to rely on temporal order rather than appearance shortcuts. Adding the motion auxiliary loss instead improves it to 68.5% ($+1.1\%$), demonstrating that explicit supervision helps the motion encoder extract motion information. Combining both losses reaches 69.0% ($+1.6\%$), verifying that the two objectives are complementary.

Table 4: Ablation study on modules. "Relation", "Context", and "Counterfactual" denote the track-to-way relation encoder, the interaction-aware contextualization, and the counterfactual temporal enhancement, respectively. $\Delta(\mathcal{J}\&\mathcal{F})$ reports the gain over the previous row.

ID	Relation	Context	Counterfactual	$\mathcal{J}\&\mathcal{F}$	$\Delta(\mathcal{J}\&\mathcal{F})$	N-Acc
I	×	×	×	65.5	–	11.1
II	✓	×	×	66.3	+0.8	21.5
III	✓	✓	×	67.9	+1.6	23.2
IV	✓	✓	✓	69.0	+1.1	23.1

Table 5: Auxiliary-loss ablation on MOVE (2-way-1-shot OS). "Motion" and "Contrastive" denote the motion auxiliary loss and the motion-aware contrastive loss, respectively.

ID	Motion	Contrastive	$\mathcal{J}\&\mathcal{F}$
I	×	×	67.4
II	×	✓	67.9
III	✓	×	68.5
IV	✓	✓	69.0



Figure 3: Qualitative comparison on MOVE between DMA (Ying et al., 2025) and ReMoM. (a) shows the interaction-dependent motions of *riding a bike* and *riding a horse*. (b) presents the temporally correlated motions: *gesture of zooming in vs. gesture of zooming out*.

4.5 Qualitative Results

Fig. 3 presents two motion-ambiguous episodes comparing our ReMoM with the baseline method DMA (Ying et al., 2025). In case (a), we showcase the interaction-dependent motions: a person riding a horse and a person riding a bike in the support videos. DMA fails to discriminate the two motions, whereas our method correctly segments the person riding a horse in the query video. This demonstrates the effectiveness of our interaction-aware contextualization module. In case (b), we show a scenario with strong temporal correlations between frames in the support set: the gestures of zooming in and zooming out. While DMA confuses the two motions due to insufficient temporal modeling, our proposed method accurately performs motion matching and segmentation. This verifies the motion understanding capability of our ReMoM.

5 Conclusion

In this work, we propose ReMoM for motion-guided few-shot video object segmentation. We introduce three complementary components to model support-defined motion relations: a track-to-way relation encoder for learning the motion relation between query tracks and support motion exemplars, a counterfactual temporal enhancement module for strengthening temporal evidence, and an interaction-aware contextualization module for resolving interaction-dependent ambiguities. Experiments on MOVE and A2D-Motion demonstrate that ReMoM consistently outperforms previous state-of-the-art baselines, validating its effectiveness in addressing the challenges of motion understanding across videos in M-FSVOS.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure: implementing methods. We used generative AI tools for the following tasks with recommended disclosure: drafting parts of this research paper, editing the research paper to improve readability, and creating or modifying scientific figures or images. We have reviewed all AI-assisted work. Specifically, we manually inspected AI-assisted code for correctness and consistency with the described methods. We also checked and revised AI-assisted text for technical accuracy, clarity, and consistency with our results. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

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 *European Conference on Computer Vision (ECCV)*, 2020.

Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14040–14049, 2021.

Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson–shiffrin memory model. In *European Conference on Computer Vision (ECCV)*, 2022.

Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Henghui Ding, Kaining Ying, Chang Liu, Shuting He, Xudong Jiang, Yu-Gang Jiang, Philip H. S. Torr, and Song Bai. MOSEv2: A more challenging dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2508.05630*, 2025.

Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Lingyi Hong, Zhongying Liu, Wenchao Chen, Chenzhi Tan, Yuang Feng, Xinyu Zhou, Pinxue Guo, Jinglun Li, Zhaoyu Chen, Shuyong Gao, Wei Zhang, and Wenqiang Zhang. LVOS: A benchmark for large-scale long-term video object segmentation. *arXiv preprint arXiv:2404.19326*, 2024.

Amin Karimi and Charalambos Poullis. DSV-LFS: Unifying LLM-driven semantic cues with visual features for robust few-shot segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4584–4594, 2025.

Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian Conference on Computer Vision (ACCV)*, 2018.

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.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Naisong Luo, Yuan Wang, Rui Sun, Guoxin Xiong, Tianzhu Zhang, and Feng Wu. Exploring the better correlation for few-shot video object segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: Unified referring video object segmentation network with a large-scale benchmark. In *European Conference on Computer Vision (ECCV)*, 2020.

Mennatullah Siam. Temporal transductive inference for few-shot video object segmentation. *International Journal of Computer Vision*, 133:4465–4482, 2025. doi: 10.1007/s11263-025-02390-x.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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 Transactions on Circuits and Systems for Video Technology*, 34(8):6699–6709, 2024.

Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Jin Wang, Bingfeng Zhang, Jian Pang, Honglong Chen, and Weifeng Liu. Rethinking prior information generation with CLIP for few-shot segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. PANet: Few-shot image semantic segmentation with prototype alignment. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Huaxin Xiao, Jiashi Feng, Guosheng Lin, Yu Liu, and Maojun Zhang. MoNet: Deep motion exploitation for video object segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Chenliang Xu, Shao-Hang Hsieh, Caiming Xiong, and Jason J. Corso. Can humans fly? action understanding with multiple classes of actors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2264–2273, 2015.

Qianxiong Xu, Lanyun Zhu, Xuanyi Liu, Guosheng Lin, Cheng Long, Ziyue Li, and Rui Zhao. Unlocking the power of SAM 2 for few-shot segmentation. In *International Conference on Machine Learning (ICML)*, 2025.

Kaining Ying, Hengrui Hu, and Henghui Ding. MOVE: Motion-guided few-shot video object segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

A Supplementary Material

A.1 Implementation details

We keep SAM 3 (Carion et al., 2025) frozen as the masklet generator and train the matching model end-to-end on episodes. The motion encoder is a 3-layer residual dilated temporal convolutional network, and the appearance encoder is a 5-layer local temporal transformer. We use a gated residual encoder ϕ for counterfactual temporal enhancement. The loss weights are $\lambda_{\mathrm{rm}}{=}1.0$, $\lambda_{\mathrm{null}}{=}0.5$, $\lambda_{\mathrm{mw}}{=}1.5$, and $\lambda_{\mathrm{cf}}{=}0.4$, with a margin $m_0{=}0.2$ and a temperature $\tau{=}0.1$. We optimize with AdamW using per-component learning rates ($3\mathrm{e}{-5}$ for the motion encoder, $1\mathrm{e}{-5}$ for the appearance encoder, and $1\mathrm{e}{-4}$ for the relation and prediction heads) and weight decay $1\mathrm{e}{-2}$. For MOVE, we train for $225{,}000$ episodes. For A2D-Motion, we train for $120{,}000$ episodes and evaluate $10{,}000$ episodes in the *2-way-1-shot* setting. For MOVE, we train on 6 GPUs with a batch size of 5 per GPU and 7500 iterations per GPU.

Masklet generation. We prompt the frozen SAM 3 model with a grid of points and propagate the prompts through each video to obtain temporally consistent candidate masklets. We then remove near-duplicate masklets using non-maximum suppression.

Motion encoder. We compute signed differences between adjacent ℓ_2-normalized object tokens, project them into a 128-dimensional space, and apply three residual dilated temporal convolutions with kernel size 3 and dilations $1, 2, 4$. We aggregate the resulting features by concatenating masked mean and max pooling statistics. For the motion auxiliary loss, we use a classification head during training only. This head computes multi-head cosine similarity between normalized query and way features and applies a learnable temperature scale.

Counterfactual temporal enhancement. We construct a counterfactual view by shuffling the query tokens along the temporal dimension while keeping the support tokens fixed as the factual anchor. This preserves appearance while disrupting temporal order. The residual encoder ϕ projects the motion difference $\boldsymbol{\delta} = \boldsymbol{r}^{\mathrm{mot}} - \boldsymbol{r}^{\mathrm{mot,cf}}$ with a bias-free two-layer MLP with GELU. We condition the projected features on the appearance relation $\boldsymbol{r}^{\mathrm{app}}$ using a gate composed of layer normalization, a linear layer, and $\tanh$: $\boldsymbol{h} = \phi(\boldsymbol{\delta}, \boldsymbol{r}^{\mathrm{app}}) = \mathrm{proj}(\boldsymbol{\delta}) \odot (1 + \mathrm{gate}(\boldsymbol{r}^{\mathrm{app}}))$. The enhanced motion relation is $\tilde{\boldsymbol{r}} = \mathrm{LN}(\boldsymbol{r}^{\mathrm{mot}} + \alpha\, \boldsymbol{h})$, where α is a non-negative learnable scalar initialized to 0.1.

A.2 Baselines

We compare ReMoM with three SAM-based baselines, all of which operate on object tokens extracted by the frozen SAM 3 model.

SAM 3 Raw. This parameter-free baseline matches each query masklet to the support prototypes using cosine similarity. We average the valid frame-level object tokens over time and apply ℓ_2 normalization to obtain one feature vector per track. We similarly average the support shots to form one prototype per way. Each query masklet is assigned to the way with the highest cosine similarity, or to the null way if this similarity is below 0.5.

SAM 3 + TimeAttn. We encode each track using a two-layer temporal transformer with eight attention heads, a feed-forward expansion factor of four, GELU activations, pre-normalization, and sinusoidal temporal position encodings. We pool the sequence using attention and aggregate the support shots into way prototypes through an attention-weighted average. A learnable prototype represents the null way. Each masklet is classified using scaled dot products between its query features and the way prototypes.

SAM 3 + SOLA. We adapt SOLA (Kim et al., 2024) to the motion-guided setting by replacing its language tokens with support object tokens and its original training loss with our $W + 1$ way classification loss.

A.3 Construction of A2D-Motion

A2D-Motion adapts the A2D action dataset (Xu et al., 2015) to the M-FSVOS setting. A2D contains $3{,}782$ videos and $11{,}926$ annotated frames. Each video is annotated with one or more of nine actions: eight regular actions and a "none" category for undefined actions.

Action folds and video-disjoint split. We partition the eight regular actions into four folds, holding out two actions for testing and using the remaining six for training in each fold. Among videos with regular action annotations, a video is assigned to the test set if any annotated frame contains a held-out action. Training videos contain only the six seen actions and no "none" annotations. The 993 videos annotated only with "none" are distributed as evenly as possible across the four folds. Each fold uses its assigned "none" videos exclusively as negative (null) queries at inference, and these videos are never used for training or as support videos. This construction keeps the training and test videos disjoint within each fold.

Table 6: A2D-Motion action folds.

Fold	Held-out actions
0	climbing, jumping
1	crawling, rolling
2	eating, flying
3	running, walking

A.4 Motion analysis

To verify that the model's decisions are driven by motion rather than by appearance, we compare three query types at inference, all applied to the query object tokens: *factual* (the original query tokens), *shuffle* (shuffling the query tokens along the temporal dimension), and *static* (replacing every query token with a single query token randomly selected from the original token sequence). As shown in Table 7, as shuffle destroys temporal order while preserving appearance and static removes temporal dynamics entirely, a drop in $\mathcal{J}\&\mathcal{F}$ under these

Table 7: Motion analysis on MOVE (2-way-1-shot OS).

Query type	$\mathcal{J}\&\mathcal{F}$
factual	69.0
shuffle	63.5
static	61.1

interventions indicates that the model exploits motion. ReMoM degrades from 69.0% (factual) to 63.5% (shuffle) and 61.1% (static), demonstrating that it relies on temporal structure. Notably, *static* drives the input of our motion encoder to zero as the encoder operates on the adjacent differences of the object tokens, because replacing every valid query token with a single query token makes all differences vanish. Therefore, the model falls back to appearance alone. Its non-trivial 61.1% reflects the dataset structure: the query video is chosen to contain the support's motion, and the object performing it often shares the object category and appearance with the support object, so appearance-only matching already succeeds on many episodes. The drop from 69.0% to 61.1% thus isolates the episodes that genuinely demand motion understanding.

A.5 Additional qualitative results

Fig. 4 and Fig. 5 provide additional qualitative comparisons between DMA and ReMoM on MOVE. These comparisons demonstrate that ReMoM better captures the motion patterns and produces fine-grained masks.

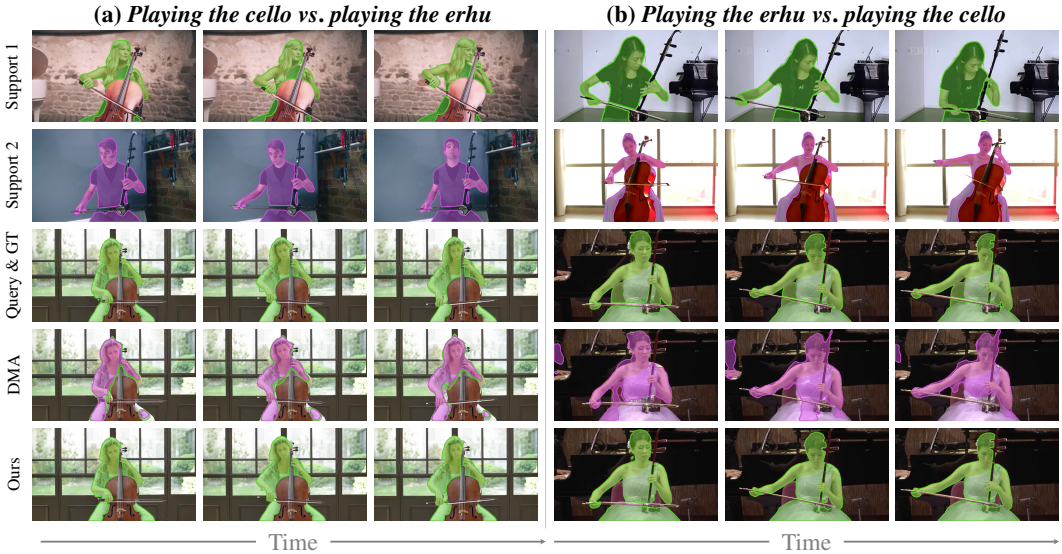

Figure 4: Additional qualitative comparisons between DMA (Ying et al., 2025) and ReMoM on MOVE: *playing the cello vs. playing the erhu* and the reverse pair.

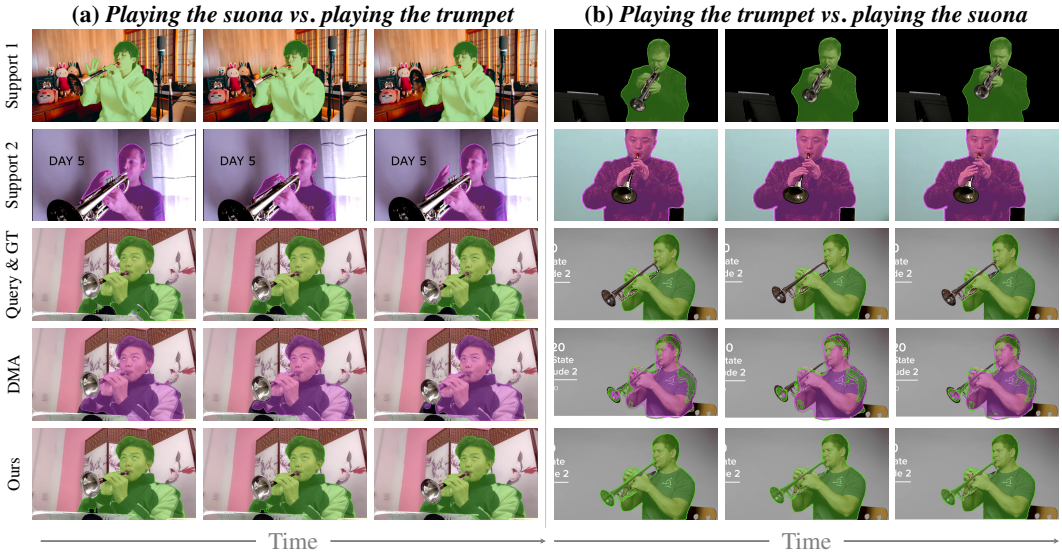

Figure 5: Additional qualitative comparisons between DMA (Ying et al., 2025) and ReMoM on MOVE: *playing the suona vs. playing the trumpet* and the reverse pair.