electronics

MDPI

Article

BiASTM-CL: Bidirectional Adaptive Spatiotemporal Modeling with Contrastive Learning for Few-Shot Action Recognition

Jing Huang * and Zijian Zhao

School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China; 2023220704097@mails.zstu.edu.cn

* Correspondence: xinjingh@zstu.edu.cn

Abstract

In few-shot action recognition (FSAR), limited annotated data and large scene variations make it difficult for models to learn stable spatial semantics and reliable temporal dynamics. As a result, spatiotemporal representations tend to be weak, and models often fail to focus on discriminative motion regions or capture frame-to-frame changes accurately. Furthermore, the insufficient fusion of local details and global context renders the learned features more susceptible to background noise and scene bias. These issues become more pronounced when background clutter is severe or when different action classes share locally similar segments, leading to unreliable support–query matching and shifted similarity distributions, which ultimately result in class confusion. To address these challenges, we propose a bidirectional adaptive spatiotemporal modeling method integrated with contrastive learning for FSAR. The method constructs attention-guided bidirectional differencing features to model inter-frame variations with semantic alignment, while suppressing background noise. It introduces a local–global interactive channel attention module to strengthen both local and global dynamic representations, and integrates dynamic distance adjustment with hard negative mining during tuple-level matching. This combination imposes contrastive constraints that enhance intra-class compactness and inter-class separability, thereby mitigating interference from cross-class similar segments. Experiments under the standard 5-way 1-shot/5-shot protocol demonstrate consistent improvements across multiple datasets, and the proposed method achieves the best performance under the 5-shot setting while remaining competitive under the 1-shot setting.

Keywords: few-shot action recognition; spatiotemporal feature extraction; contrastive learning; tuple-level matching

Check for updates

Academic Editor: George A. Tsihrintzis

Received: 5 March 2026
Revised: 2 April 2026
Accepted: 3 April 2026
Published: 14 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Video action recognition has made significant progress in practical scenarios such as intelligent security [1] and anomaly detection [2]. Recent applications also span factory safety monitoring [3,4], fall detection in healthcare [5], surgical workflow understanding [6], and sports video analytics [7]. It relies heavily on large-scale labeled data, thus limiting its applicability in data-scarce scenarios. Consequently, few-shot action recognition (FSAR) has become an active research topic, aiming to recognize novel action categories with only a few labeled samples. Compared with image recognition, action recognition needs to handle higher-dimensional data structures, and therefore faces two key challenges: temporal continuity and spatial variability. Spatiotemporal feature modeling has improved significantly in recent years, laying the foundation for FSAR [8–13]. Most FSAR methods

adopt a metric-learning paradigm by learning a shared embedding space and measuring the similarity between the support set and the query set to enable cross-class transfer. Representative FSAR studies have mainly focused on temporal alignment, structured matching, and metric learning, while recent studies have continued to improve feature transferability and spatiotemporal representation learning in FSAR [14–16].

Existing methods achieve promising recognition performance and continue to enhance spatiotemporal modeling, yet they still face two major challenges. First, actions typically evolve over time and exhibit complex spatial structures, making it difficult to characterize complete dynamics when relying only on a few key frames or local details. Second, some actions appear visually similar but are semantically different; such differences are often manifested in only a few key frames or fine-grained regions. Traditional methods using fixed metrics may fail to capture these subtle differences robustly, leading to inter-class confusion and weakened discrimination. To address these issues, the model should possess stronger temporal awareness and better focus on key motion changes. It should jointly model local details and global pose, while maintaining reliable separation of similar sub-sequences in the metric space.

Motivated by these observations, we propose a new few-shot action recognition method termed Bidirectional Adaptive Spatiotemporal Modeling with Contrastive Learning (BiASTM-CL). The proposed method follows two synergistic ideas. First, we introduce inter-frame attention into differencing-based modeling, which stabilizes differenced features via semantic alignment and forms bidirectional dynamic representations. Second, we incorporate local–global interactive channel attention on top of the differenced features to enhance salient cues and suppress redundant information, and further impose contrastive constraints with dynamic distance adjustment during tuple-level matching. This design improves intra-class compactness and inter-class separability, thereby reducing mismatches caused by similar sub-sequences.

To summarize, the major contributions of this work are as follows:

- We design an Attention-Guided Inter-Frame Differencing Mechanism (AGID), which leverages cross-frame attention to drive bidirectional difference modeling and construct forward–backward temporal difference sequences, thereby capturing critical motion dynamics in actions.
- We propose a Local–Global Interactive Channel Attention Module (LGICA) that simultaneously models local details and global dependencies along the channel dimension. Driven by cross-correlation, it adaptively generates attention weights and collaborates with spatial attention to achieve deep integration of fine-grained and global semantics.
- We develop a Dynamic Distance Adjustment Contrastive Learning Module (DACM). It adaptively adjusts intra-class compactness and inter-class separability to refine similarity scores between support and query samples, effectively suppressing interference from similar sub-sequences and background clutter.

Problem Formulation

Few-shot action recognition is formulated under the episodic meta-learning setting. Let the dataset be divided into a training set D_{train} and a test set D_{test}, where the corresponding class sets are disjoint, i.e.,

$$\mathcal{C}_{\text{train}} \cap \mathcal{C}_{\text{test}} = \varnothing.$$

In each episode, an N-way K-shot task is constructed by sampling N action classes, with K labeled videos per class forming the support set

$$\mathcal{S} = \{(x_i^c, y^c) \mid c = 1, \ldots, N,\ i = 1, \ldots, K\},$$

and the corresponding query videos forming the query set

$$\mathcal{Q} = \{(x_j^c, y^c) \mid c = 1, \ldots, N,\ j = 1, \ldots, M\}.$$

Given a query video $x_q \in \mathcal{Q}$, the goal is to predict its class label by matching it against the support classes:

$$\hat{y}_q = \arg \max_{c \in \{1,\ldots,N\}} \mathrm{Sim}\big(\Phi(x_q), \Phi(\mathcal{S}^c)\big),$$

where $\Phi(\cdot)$ denotes the learned video representation and $\mathrm{Sim}(\cdot,\cdot)$ denotes the support–query similarity function.

The main challenge is to learn a representation and matching function that remain reliable under limited supervision, temporal variation, and cross-class similar sub-sequences. Ideally, the similarity between a query and its target class should be higher than that to non-target classes, i.e.,

$$\mathrm{Sim}\big(\Phi(x_q), \Phi(\mathcal{S}^{y_q})\big) > \mathrm{Sim}\big(\Phi(x_q), \Phi(\mathcal{S}^c)\big), \quad \forall c \neq y_q.$$

Motivated by this objective, our method focuses on temporally informative differencing-based motion modeling, local–global feature refinement, and interference-aware similarity learning.

2. Related Works

2.1. Few-Shot Learning

Few-shot learning (FSL) aims to train deep learning models with strong generalization ability using only a small number of labeled samples. It is widely applied in data-scarce scenarios such as action recognition, image classification, and medical diagnosis. Most FSL approaches can be broadly categorized into three paradigms: memory-based, optimization-based, and metric-based methods.

Memory-based methods introduce an external memory to store and reuse knowledge across tasks, alleviating the lack of labeled samples in the current task. MM-Net [17] integrates a differentiable memory module with a matching-network architecture, enabling the model to quickly retrieve task-relevant representations learned from previous episodes when adapting to new tasks.

Optimization-based methods design learning procedures or model structures tailored to the few-shot setting, allowing rapid parameter adaptation from only a few labeled examples. MetaNODE [18] treats class prototypes as optimizable variables and models their evolution with neural ordinary differential equations, so that class centers can be updated efficiently and flexibly.

Metric-based methods perform matching between support and query samples by learning a discriminative embedding space together with an appropriate distance metric. Prototypical Networks [19] construct a prototype vector for each class and embed query instances into the same feature space. Classification is then performed by measuring Euclidean distances between the query features and class prototypes, which facilitates recognition of unseen classes.

2.2. Few-Shot Action Recognition

Few-shot action recognition (FSAR) is an important few-shot learning (FSL) task in video understanding. Its goal is to recognize novel action classes when only a few labeled support examples are available. Because video annotation is expensive and class distributions are often unbalanced, FSAR has attracted increasing attention in recent years. Most existing methods adopt metric learning, where a shared feature space and a similarity

measure are learned for support–query matching in episodic training and knowledge transfer to novel classes [20,21].

Temporal modeling is the core of FSAR, and many existing methods improve support–query matching through temporal alignment or structured spatiotemporal modeling. TRX [22] uses tuple-level matching to reduce the impact of temporal shifts and speed differences. TRX-IDP [23] further introduces an image difference pyramid to capture finer temporal changes. HyRSM [24] and TA2N [25] adopt staged alignment to reduce mismatches caused by variations in action duration and tempo. OTAM [26] employs dynamic time warping (DTW) for sequence alignment to achieve more structured temporal matching. TEAM [27] explores an alignment-free matching paradigm without requiring explicit temporal correspondence. FTAN [28] studies frame-to-frame alignment and temporal cyclic shift, thereby improving robustness to temporal inconsistency.

Improving feature quality is also important for reliable similarity computation. STRM [15] combines local patch cues and global frame context to learn more discriminative features with limited labels. MoLo [16] uses multi-scale contrastive learning to connect global semantics with local motion cues, which improves class separation. FICAMA [29] uses adaptive metric aggregation to make similarity scores more stable. TSA-MLT [30] adds task-specific alignment and multi-level temporal feature learning to describe complex action dynamics better.

Although the aforementioned methods have achieved remarkable progress in FSAR, several challenges remain. On the one hand, existing spatio-temporal representation learning often struggles to jointly preserve fine-grained local cues and capture long-range global dynamics, especially under strong background clutter or pronounced long-term temporal evolution. On the other hand, suppressing interference from cross-class similar segments is still inadequate: highly similar yet semantically different sub-sequences, as well as background cues, may yield misleading high-similarity responses, thereby distorting support–query similarity distributions and inducing mismatches during fine-grained comparison. These observations highlight the need for stronger temporal change modeling, more adaptive local–global feature integration, and interference-aware similarity learning with adaptive distance calibration.

3. Method

To enhance the modeling capability and discriminative performance of few-shot action recognition in complex spatio-temporal scenarios, we propose a bidirectional adaptive spatiotemporal modeling method integrated with contrastive learning. As illustrated in Figure 1, our method follows the episodic meta-learning setting defined in Section Problem Formulation, using subsequence-level units as the basic modeling elements. The Attention-Guided Inter-Frame Differencing Mechanism (AGID) and Local–Global Interactive Channel Attention Module (LGICA) collaboratively capture global and local dynamics from bidirectional differential features. Furthermore, the DACM module refines mutual similarity estimation and contrastive learning between support and query sets, thereby strengthening the model's ability to capture complex temporal dynamics and distinguish fine-grained action categories.

For each input video sequence, we follow the sparse temporal sampling strategy of TSN [31] to sample T frames from each video. The support set is represented as $S = \{S_1, S_2, \ldots, S_T\}$, and the query set as $Q = \{Q_1, Q_2, \ldots, Q_T\}$. Each frame in both the support and query sets is processed by a feature extraction network to obtain frame-level features.

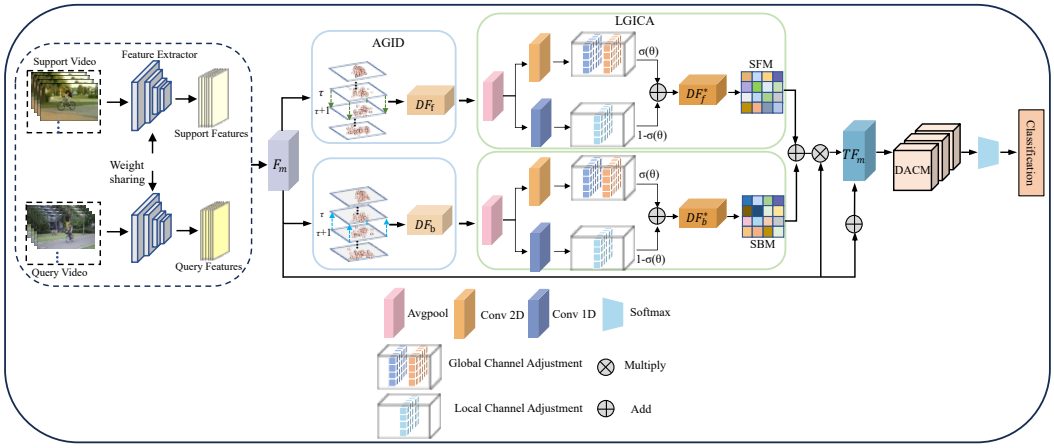


Figure 1. Overall framework of our method.

3.1. Attention-Guided Inter-Frame Differencing Mechanism.

To better capture subject motion during action execution and reduce interference from irrelevant regions in traditional frame differencing models, we propose an Attention-Guided Inter-Frame Differencing Mechanism (AGID). As shown in Figure 2, AGID extends traditional frame differencing by incorporating inter-frame attention, enabling the differencing computation to be semantically guided, thereby focusing more on action-relevant regions.

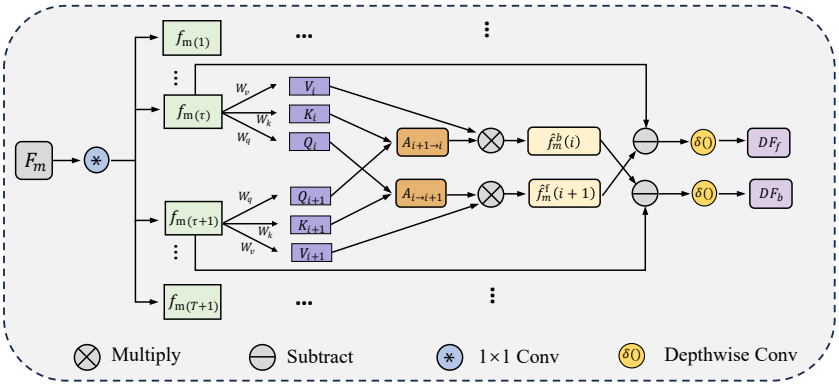


Figure 2. Illustration of the Attention-Guided Inter-Frame Differencing Mechanism.

In particular, AGID models the semantic relationship between the current and neighboring frames using an inter-frame attention mechanism, thereby enhancing the semantic consistency between adjacent frame features. Given the input feature sequence

$$F_m = \left[f_m(1), f_m(2), \ldots, f_m(T)\right], \tag{1}$$

where $f_m(\tau)$ denotes the feature representation of the τ-th frame and T is the sequence length, AGID models the semantic relationship between adjacent frames using two guiding branches: the forward guidance branch generates queries from the current frame features and keys and values from the neighboring frame features. By calculating the inter-frame attention weights for $\tau \rightarrow \tau + 1$, the neighboring frame features are weighted and reconstructed, yielding the semantically enhanced next-frame feature $\hat{f}_m(\tau + 1)$. The backward guidance branch symmetrically models attention from $\tau + 1 \rightarrow \tau$, producing a semantically enhanced version of the current frame feature, denoted as $\hat{f}_m(\tau)$.

To alleviate unstable differencing responses caused by minor spatial inconsistency between adjacent frames, AGID introduces a depthwise separable convolution operator $\delta(\cdot)$ after differencing as a lightweight local filtering step. Rather than performing explicit

spatial alignment, this operation acts as a local filtering step that aggregates neighborhood responses and suppresses isolated noisy activations, thereby improving the stability of the differenced features. The forward differencing describes the dynamic change from $\tau \to \tau + 1$, while the backward differencing captures the dynamic change from $\tau + 1 \to \tau$, defined as follows:

$$DF_f = \delta\big(\hat{f}_m(\tau + 1) - f_m(\tau)\big), \qquad DF_b = \delta\big(\hat{f}_m(\tau) - f_m(\tau + 1)\big). \tag{2}$$

where DF_f and DF_b represent the attention-guided forward- and backward-differenced features, respectively. As complementary bidirectional dynamic cues, they are fed into later spatiotemporal modules for fusion. This helps the model better capture action changes over time and semantic relations between frames.

3.2. Local–Global Interactive Channel Attention Module

Different channels respond to motion in different ways. Some channels react more to short, local changes, while others focus more on global action semantics over time. Based on this, we introduce the Local–Global Interactive Channel Attention (LGICA) module on top of bidirectional difference features. Unlike a simple parallel combination of local and global branches, LGICA explicitly models the interaction between local channel dependencies and global channel context, thereby providing more discriminative channel guidance for later spatial attention.

As shown in Figure 3, LGICA takes the forward differenced feature $DF_f \in \mathbb{R}^{C \times H \times W}$ as input, where C is the number of channels and H, W are the spatial sizes. We first apply global average pooling to compress spatial information and obtain a channel-wise descriptor vector G_n^f. This vector reflects the overall response strength of each channel in the current differenced feature while preserving channel discrimination.

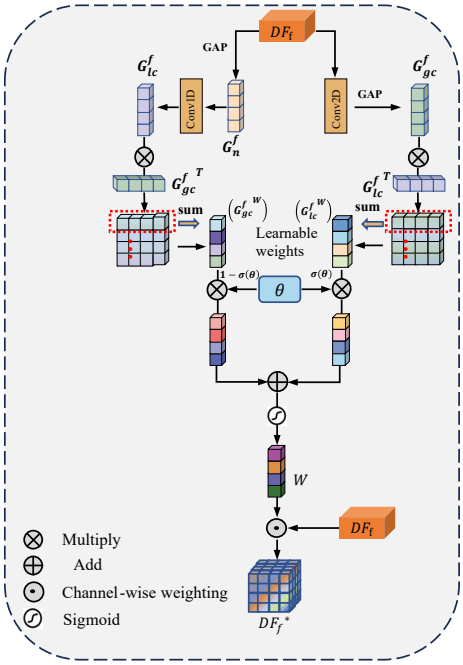


Figure 3. Illustration of the Local–Global Interactive Channel Attention Module.

LGICA constructs two complementary branches for local and global channel modeling. In the local branch, we apply a 1D convolution along the channel dimension to the pooled channel descriptor G_n^f to aggregate nearby channels and produce G_{lc}^f. The global branch applies a 2D convolution directly to the differenced feature DF_f to capture broader channel

context, and the resulting feature is further compressed by global average pooling to obtain the global channel representation G_{gc}^{f}.

Importantly, the effectiveness of LGICA does not come from directly concatenating or simply summing these two branches. Instead, the interplay between the local and global channel representations is explicitly modeled via cross-correlation:

$$A = G_{gc}^{f} \cdot \left(G_{lc}^{f}\right)^{\top}. \tag{3}$$

where A measures the degree of alignment between channels based on the global and local channel representations. By jointly leveraging A and its transpose $A^{\top}$, we perform row-wise summation on A to extract row-vector information and on $A^{\top}$ to extract column-vector information, which are then used to derive the global channel attention weights $\left(G_{gc}^{f}\right)^{W}$ and the local channel attention weights $\left(G_{lc}^{f}\right)^{W}$. During attention fusion, a learnable factor θ is introduced and combined with the sigmoid activation $\sigma(\cdot)$ to adaptively weight the two channel attention terms, producing the final fused channel weights W:

$$W = \sigma\left(\sigma(\theta) \times \sigma\left(\left(G_{gc}^{f}\right)^{W}\right) + (1 - \sigma(\theta)) \times \sigma\left(\left(G_{lc}^{f}\right)^{W}\right)\right). \tag{4}$$

In this way, LGICA can recalibrate channel responses more effectively by jointly considering short-range local motion variations and broader global action semantics, rather than relying on a simple dual-branch aggregation scheme. The fused channel weights are then utilized to recalibrate the input forward differential feature, i.e., to adjust the channel-wise responses of the feature map:

$$DF_{f}^{*} = W \odot DF_{f}, \tag{5}$$

where DF_{f}^{*} denotes the channel-enhanced forward-differenced feature, and $\odot$ denotes element-wise multiplication, which enhances discriminative channel responses that are closely related to action variations.

The channel-enhanced forward-differenced feature is further used for spatial attention modeling. Applying a Softmax operation over the spatial dimension yields the forward spatial attention map SFM, which captures the spatial saliency distribution of forward motion variations. The backward spatial attention map SBM is obtained in the same manner from the backward-differenced feature; due to the symmetric design, we omit the repeated description.

Finally, the forward and backward spatial attention maps jointly participate in spatio-temporal feature fusion to produce the enhanced representation:

$$TF_{m} = \lambda\left(SFM + SBM\right)\delta(F_{m}) + F_{m}, \tag{6}$$

where λ is a learnable coefficient that balances the contributions of the original and enhanced features.

3.3. Dynamic Distance Adjustment Contrastive Learning Module

In few-shot action recognition, different classes often contain numerous subsequences that are locally similar in appearance yet semantically distinct. Such interference is particularly prominent during tuple-level matching, leading the model to misclassify local similarity as class consistency. To address this issue, we develop a Dynamic Distance Adjustment Contrastive Learning Module (DACM) on top of the TRX framework. By explicitly introducing intra-class and inter-class structural constraints together with contrastive learning, DACM alleviates confusion caused by cross-class similar sub-sequences.

Specifically, as illustrated in Figure 4, after frame-level feature extraction and temporal grouping, we represent a query video as a set of query tuples: $\{TF_{qt_1}, TF_{qt_2}, \ldots, TF_{qt_L}\}$, where L denotes the number of query tuples. Each tuple $TF_{qt_i} \in \mathbb{R}^{\omega \times d}$ corresponds to a local temporal segment composed of ω frames, and d is the feature dimension. Similarly, let $TF_{st}^{c,j}$ denote the j-th support tuple from class c. We first compute the base tuple-level distance between the i-th query tuple and the j-th support tuple from class c:

$$D_{i,j}^{c} = d\left(TF_{qt_i}, TF_{st}^{c,j}\right), \tag{7}$$

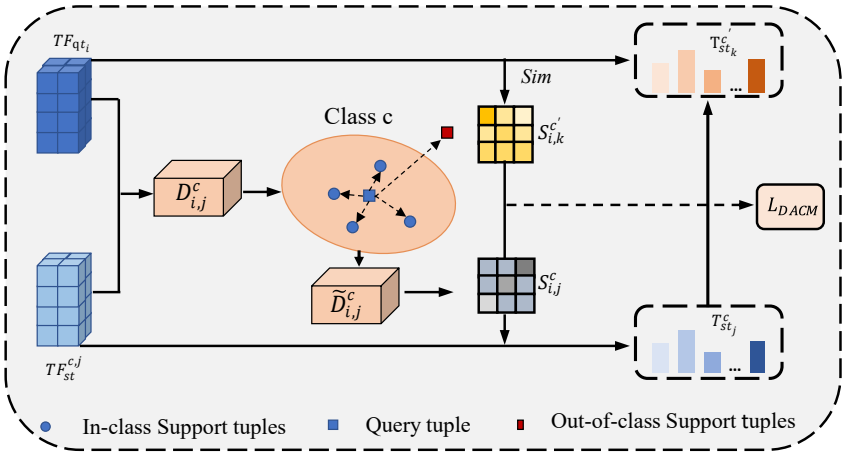


Figure 4. Illustration of the Dynamic Distance Adjustment Contrastive Learning Module. Colored squares denote similarity responses, while blue and orange bars denote in-class and out-of-class support tuples, respectively.

To incorporate class-level structural information, we further define two class-aware reference distances for the query tuple TF_{qt_i} with respect to class c. The first is the intra-class mean distance at the tuple level, where S_c denotes the set of indices of support tuples extracted from the support video(s) of class c. This term measures the average distance from the query tuple to all support tuples in class c, rather than to other videos of the same class. Therefore, even in the N-way 1-shot setting, $\bar{D}_{\text{intra}}^{c}(i)$ remains well-defined, since multiple temporal tuples can still be extracted from the single support video of class c:

$$\bar{D}_{\text{intra}}^{c}(i) = \frac{1}{|S_c|} \sum_{j \in S_c} D_{i,j}^{c}, \tag{8}$$

The second is the inter-class reference distance, which measures the smallest distance from the query tuple to support tuples belonging to any non-target class:

$$\bar{D}_{\text{inter}}^{c}(i) = \min_{c' \neq c} \min_{k \in S_{c'}} D_{i,k}^{c'}, \tag{9}$$

Based on these two quantities, we define the dynamically adjusted distance for the j-th support tuple in class c as

$$\tilde{D}_{i,j}^{c} = D_{i,j}^{c} + \alpha\, \bar{D}_{\text{intra}}^{c}(i) - \beta\, \bar{D}_{\text{inter}}^{c}(i), \tag{10}$$

where α and β are hyperparameters controlling the contributions of the intra-class tuple aggregation term and the inter-class separation term, respectively. We use the class-wise mean for the intra-class term to reflect the overall consistency of the target class, while the minimum cross-class distance is used to capture the most confusing non-target competitor. In this way, the tuple-level distance is calibrated not only by the candidate support tuple

itself, but also by the overall structural relationship between the query tuple and the support classes.

Subsequently, we define the target-class similarity as $S_{i,j}^{c} = -\widetilde{D}_{i,j}^{c}$, so that a smaller adjusted distance corresponds to a larger similarity. The support tuple with the highest similarity in the target class is then selected as the positive sample for contrastive learning. Accordingly, we construct the positive-class support representation $T_{st_i}^{c}$, where $T_{st_i}^{c}$ denotes the support representation from the target class that best matches TF_{qt_i}.

Interference from similar sub-sequences often arises from negative-class segments that appear even more similar than positive samples. Therefore, DACM explicitly mines highly similar negative samples. For any non-target class $c' \neq c$, we compute the cross-class similarity between the i-th query tuple and the k-th support tuple from class c' as

$$S_{i,k}^{c'} = \left\langle \phi_{\text{cls}}\left(TF_{qt_i}\right), \phi_{\text{cls}}\left(TF_{st_k}^{c'}\right) \right\rangle, \tag{11}$$

where $\phi_{\text{cls}}(\cdot)$ denotes the projection head and $\langle \cdot, \cdot \rangle$ is the inner product. We treat the support tuple as a highly similar negative sample if

$$S_{i,k}^{c'} > \max_{j \in S^{c}} S_{i,j}^{c}. \tag{12}$$

When the above condition holds, we construct the negative-class support representation $T_{st_k}^{c'}$, where $T_{st_k}^{c'}$ denotes the highly similar support representation from the non-target class c'. It explicitly constrains cross-class interference that is "more similar than the positive class," encouraging the model to enlarge the discriminative margin on the most confusing sub-sequences.

Based on the constructed positive-class support representation $T_{st_i}^{c}$ and the negative-class support representation $T_{st_k}^{c'}$, DACM is optimized with a contrastive learning loss over the query tuples:

$$\mathcal{L}_{\text{DACM}} = -\frac{1}{L}\sum_{i=1}^{L} \log \frac{\max_{j \in S_c}\left(S_{i,j}^{c}\right)}{\max_{j \in S_c}\left(S_{i,j}^{c}\right) + \sum_{c' \neq c} \max_{k \in S_{c'}}\left(S_{i,k}^{c'}\right)}. \tag{13}$$

This loss encourages higher matching scores between query tuples and the positive class while suppressing matching responses to highly similar negatives, alleviating the impact of cross-class similar sub-sequences on recognition accuracy. Finally, the DACM loss is jointly optimized with the original TRX loss, and the overall objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{TRX}} + \mu \cdot \mathcal{L}_{\text{DACM}}, \tag{14}$$

where μ is a loss weighting coefficient that balances the relative contributions of the classification objective and the contrastive constraint during training.

4. Experiments

4.1. Dataset

We conduct experiments on three representative public action recognition datasets to evaluate the proposed method. UCF101 [32] has 101 action classes with 13,320 videos. The scenes are relatively clean, and the classes are easier to separate, so it is a common benchmark for few-shot action recognition. HMDB51 [33] includes 51 daily-action categories with approximately 6766 videos collected from real-world footage. Due to frequent blur and occlusions, it is widely used to assess model robustness. Something-Something V2 [34] consists of 174 fine-grained action categories dominated by hand–object interactions, cov-

ering over 168,000 short clips. It emphasizes temporal dependencies and subtle semantic differences, and is widely adopted for action recognition and video understanding studies. Figure 5 presents example action instances from the three datasets.

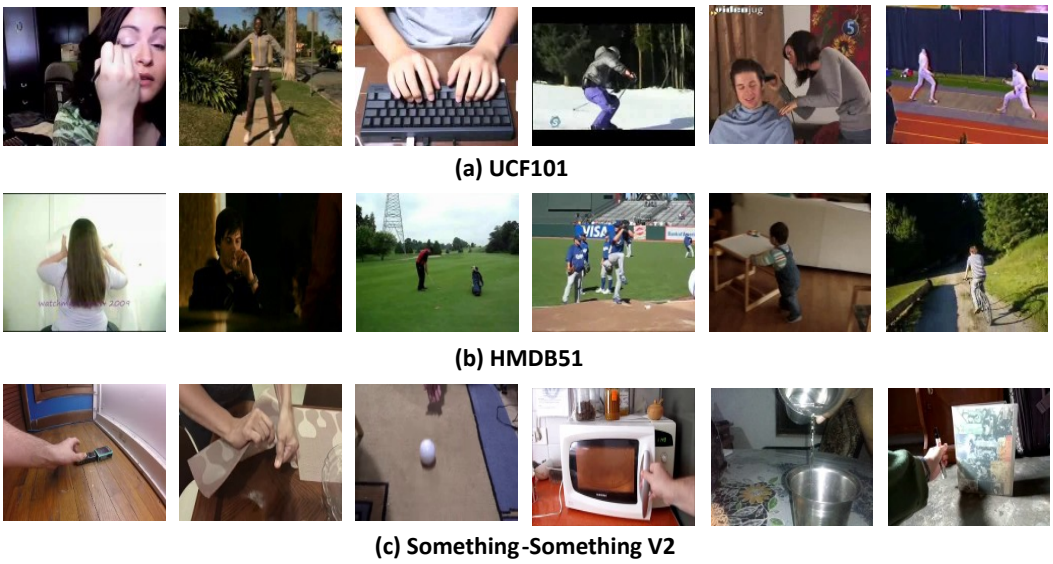


Figure 5. Example category images from three representative few-shot action recognition datasets: (**a**) UCF101, (**b**) HMDB51, and (**c**) Something-Something V2.

4.2. Experimental Setup

To ensure fair comparison, we use ResNet-50 as the feature extractor. All experiments are implemented using PyTorch v2.3.1 with CUDA v11.8 on an NVIDIA GeForce RTX 5090 D GPU. For each video, we uniformly sample $T = 8$ segments and resize each frame to 224×224. We set the learning rate to 2×10^{-4} and the batch size to 16, with $\alpha = \beta = 0.3$ and the loss weight $\mu = 0.2$. We evaluate our method on UCF101, HMDB51, and Something-Something V2, focusing on the 5-way 1-shot and 5-way 5-shot few-shot action recognition settings. To reduce the influence of randomness and improve reproducibility, all experiments are repeated over three independent runs with different random seeds, and the reported results are given as mean accuracy $\pm$ standard deviation. During these runs, the network architecture, input resolution, sampled frame number, and optimization settings remain unchanged, and only the random seed is varied.

4.3. Ablation Experiments

We conduct ablation studies on AGID, LGICA, and DACM using two benchmark datasets: UCF101 and HMDB51. We use TRX as the baseline. We evaluate each module alone, and we also evaluate the full model with all modules. These experiments examine how each module affects temporal modeling and spatiotemporal feature learning.

4.3.1. Validation of AGID and LGICA

Table 1 reports the ablation results for AGID and LGICA. Both modules are individually effective, with each bringing clear improvements over the baseline. AGID mainly enhances temporal difference modeling, whereas LGICA refines channel-level feature representation. Although the marginal gain from combining the two modules is relatively modest in some settings, their joint use still achieves the best overall results across both datasets, indicating a complementary but not strongly additive relationship.

Table 1. Ablation study of AGID and LGICA (accuracy, %).

Method	UCF101		HMDB51	
	1-Shot	**5-Shot**	**1-Shot**	**5-Shot**
Baseline (TRX)	77.3	96.1	52.9	75.6
+ AGID	83.2	96.9	57.4	76.8
+ LGICA	83.5	97.0	57.5	77.2
+ AGID + LGICA	84.0	97.2	57.6	78.0

4.3.2. DACM Validity Verification

Table 2 reports the ablation results for DACM. The results show that DACM improves discrimination between confusing action classes by reducing the influence of cross-class similar sub-sequences during tuple-level matching. When DACM is combined with AGID, the performance is further improved, indicating that reliable temporal difference modeling and interference-aware similarity learning are complementary.

Table 2. Ablation study of DACM (accuracy, %).

Method	UCF101		HMDB51	
	1-Shot	**5-Shot**	**1-Shot**	**5-Shot**
Baseline (TRX)	77.3	96.1	52.9	75.6
+ DACM	83.0	96.5	56.8	76.9
+ AGID + DACM	84.0	97.2	57.6	78.1
+ AGID + LGICA + DACM	84.3	97.4	57.8	78.8

4.3.3. Validation of the δ Operation in AGID

To further validate the effectiveness of the δ operation in AGID, we compare the full AGID with a variant without δ while keeping the inter-frame attention guidance unchanged. The results are reported in Table 3.

Table 3. Ablation study of the δ operation in AGID (accuracy, %).

Method	UCF101		HMDB51	
	1-Shot	**5-Shot**	**1-Shot**	**5-Shot**
Baseline (TRX)	77.3	96.1	52.9	75.6
+ AGID without δ	82.9	96.7	56.9	76.5
+ Full AGID	83.2	96.9	57.4	76.8

As shown in Table 3, removing the δ operation leads to lower performance across all settings. This indicates that δ contributes positively to AGID by improving the stability of the differenced features. Therefore, δ is better understood as a lightweight local filtering step rather than an explicit motion compensation mechanism.

4.3.4. DACM Hyperparameter Analysis

To evaluate the impact of the intra-class compactness factor α and the inter-class separability factor β in DACM, we conduct a hyperparameter analysis under the baseline (TRX) + DACM setting, and the results are reported in Table 4. We observe that moderately increasing α and β generally improves performance, with the best result achieved at $\alpha = 0.3$ and $\beta = 0.3$.

Table 4. Hyperparameter analysis with DACM only (accuracy, %).

α	β	UCF101		HMDB51	
		1-Shot	**5-Shot**	**1-Shot**	**5-Shot**
–	–	77.3 (Baseline)	96.1	52.9	75.6
0.1	0	80.5	96.2	54.1	76.0
0.3	0	81.7	96.3	54.7	76.5
0.3	0.1	82.4	96.3	56.2	76.5
0.3	0.3	83.0	96.5	56.8	76.9
0.5	0.3	82.9	96.3	57.0	76.7
0.5	0.5	82.4	96.4	56.6	76.3

4.3.5. DACM Component Ablation

To further analyze the contribution of each component in DACM, we separate it into two parts: dynamic distance adjustment (Equation (10)) and the contrastive constraint (Equation (13)). We conduct ablation experiments under the same training and evaluation settings as the TRX baseline, and report the results in Table 5. Enabling only dynamic distance adjustment calibrates tuple-level distances by incorporating intra-class compactness and inter-class separation, which helps reduce matching bias caused by cross-class similar sub-sequences. Enabling only the contrastive constraint focuses on hard negatives and strengthens separation between classes, further suppressing interference from highly similar cross-class segments. When both components are enabled, DACM achieves the best performance on both datasets under 1-shot and 5-shot settings, suggesting that distance-level calibration and contrastive objectives complement each other.

Table 5. Ablation study of DACM components. Distance Adjustment indicates that only the dynamic distance adjustment in Equation (10) is enabled, Contrastive Constraint indicates that only the contrastive constraint in Equation (13) is enabled, and DACM indicates that both components are enabled. Except for the listed components, all other training and evaluation settings are kept identical to the TRX baseline. ✓/× denote whether a component is enabled/disabled, respectively (accuracy, %).

Distance Adjustment (Equation (10))	**Contrastive Constraint (Equation (13))**	**UCF101**		**HMDB51**	
		1-Shot	**5-Shot**	**1-Shot**	**5-Shot**
×	×	77.3	96.1	52.9	75.6
✓	×	81.5	96.2	54.8	76.2
×	✓	81.0	96.2	55.6	76.1
✓	✓	83.0	96.5	56.8	76.9

4.3.6. Analysis of the Impact of the Number of Sampled Frames

To examine how the number of sampled frames affects performance, we conduct experiments on the UCF101 dataset under a 5-way 1-shot setting. Using the same segment-based sampling protocol as in Section 3, we uniformly sample 2 to 10 frames from each video and feed them into the model for evaluation. As shown in Figure 6, the classification accuracy generally increases as more frames are provided. The performance improves steadily from two to eight frames, reaching its peak at eight frames. When the number of frames is further increased (e.g., to 9 or 10), the accuracy slightly drops, suggesting that an appropriate frame count (e.g., 8) allows the model to better exploit spatiotemporal information for optimal performance.

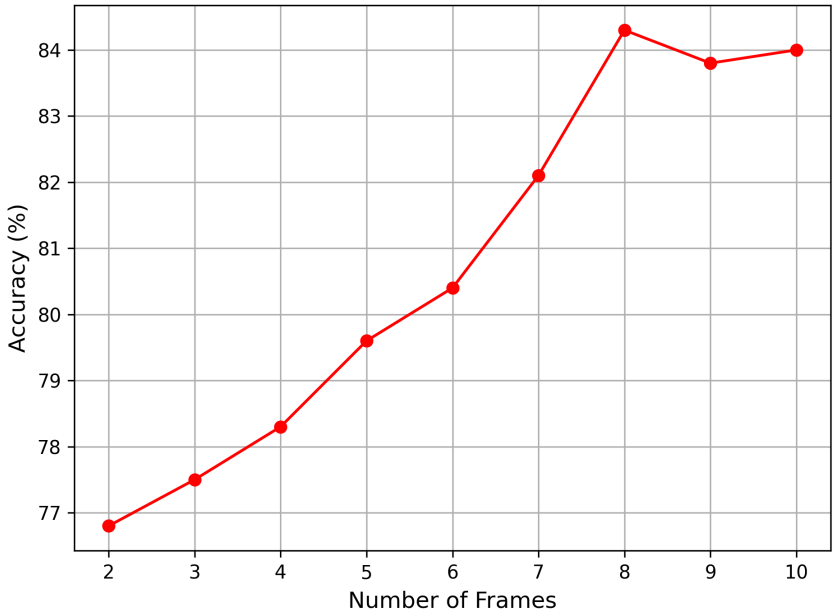


Figure 6. Impact of different numbers of input frames on UCF101.

4.4. Comparative Experiments

4.4.1. Comparison Between AGID and Conventional Frame Differencing

To validate the contribution of AGID, we compare it with conventional frame differencing under the same training configuration. Conventional frame differencing directly computes the differences between adjacent frame features. AGID incorporates inter-frame attention as semantic guidance, thereby suppressing background noise and improving the consistency and discriminability of the differenced features. As shown in Table 6, adding conventional frame differencing to the baseline improves accuracy on both datasets, confirming the usefulness of modeling frame-to-frame changes. When we replace it with AGID under the same setting, accuracy increases again, indicating that inter-frame semantic guidance yields more stable and informative motion representations for recognition.

Table 6. Comparison between AGID and conventional frame differencing (accuracy, %).

Method	UCF101		HMDB51	
	1-Shot	**5-Shot**	**1-Shot**	**5-Shot**
Baseline (TRX)	77.3	96.1	52.9	75.6
+ Difference (naïve)	82.1	96.3	54.8	76.0
+ AGID	83.2	96.9	57.4	76.8

4.4.2. Comparison of LGICA with Other Attention Mechanisms

To evaluate whether LGICA is more effective than existing attention modules, we keep all other training and evaluation settings unchanged and replace LGICA with SE, CBAM, and ECA for comparison. As shown in Table 7, LGICA achieves the best performance on both UCF101 and HMDB51 under the 1-shot/5-shot settings, consistently outperforming SE, CBAM, and ECA.

Specifically, SE relies on global-average-pooled channel statistics and a bottleneck structure to generate channel weights. Due to the strong compression, it may weaken fine-grained dynamic cues, leading to limited gains in the 1-shot setting and even a slight performance drop. CBAM introduces an additional spatial branch, but its channel and spatial attention are typically applied sequentially. Their interaction is relatively weak, and the multi-scale modeling of channel relationships is also limited. ECA generates channel

weights via local cross-channel interactions without dimensionality reduction. It yields more stable improvements on both datasets, suggesting that preserving channel details can be beneficial. However, it mainly focuses on local channel neighborhoods and lacks explicit interaction modeling between local channel dependencies and global channel context. In contrast, LGICA explicitly captures the interaction between local channel relations and global channel information through cross-correlation, thereby providing more discriminative channel guidance for highlighting action-relevant features and suppressing redundant responses.

Table 7. Replacement comparison of LGICA with SE/CBAM/ECA (accuracy, %).

Method	UCF101		HMDB51	
	1-Shot	5-Shot	1-Shot	5-Shot
Baseline (TRX)	77.3	96.1	52.9	75.6
+ SE	76.8	96.3	52.5	75.7
+ CBAM	79.4	96.6	53.7	76.2
+ ECA	81.2	96.7	55.2	76.9
+ LGICA	83.5	97.0	57.5	77.2

4.4.3. Comparison with Representative Methods

Table 8 compares BiASTM-CL with representative few-shot action recognition methods on UCF101, HMDB51, and SSV2 under the standard 5-way 1-shot and 5-shot settings. For BiASTM-CL, the results are reported as mean accuracy together with the standard deviation. Under the 5-shot setting, BiASTM-CL achieves 97.4% $\pm$ 0.2% on UCF101, 78.8% $\pm$ 0.3% on HMDB51, and 61.7% $\pm$ 0.4% on SSV2. Under the 1-shot setting, it achieves 84.3% $\pm$ 0.4%, 57.8% $\pm$ 0.5%, and 38.8% $\pm$ 0.6%, respectively. These results show that BiASTM-CL achieves the best performance on all three datasets under the 5-shot setting while remaining competitive under the 1-shot setting. Overall, these results demonstrate the effectiveness and statistical stability of BiASTM-CL.

Table 8. Comparison with representative few-shot action recognition methods. The best results are highlighted in bold. * denotes TRX and its subsequent variants. For our proposed BiASTM-CL, the results are reported as mean $\pm$ standard deviation.

Method	UCF101		HMDB51		SSV2	
	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
OTAM [26]	79.9	88.9	54.5	66.1	36.4	48.0
TA2N [25]	81.9	95.1	59.7	73.9	–	–
MoLo [16]	86.0	95.5	60.6	77.3	41.9	56.2
HyRSM [24]	83.9	94.7	60.3	76.0	40.6	56.1
HCL [13]	82.5	93.9	59.1	76.3	38.7	55.4
SloshNet [35]	86.0	97.1	59.4	77.5	–	–
TEAM [27]	**87.2**	96.2	**62.6**	78.4	**43.8**	58.2
TRX * [22]	77.3	96.1	52.9	75.6	36.0	59.1
TRX-IDP * [23]	–	96.3	–	76.5	–	–
STRM * [15]	79.2	96.9	54.1	77.3	37.1	55.3
TSA-MLT * [30]	80.6	97.1	57.9	78.2	37.2	55.3
BiASTM-CL * (ours)	84.3 $\pm$ 0.4	**97.4 $\pm$ 0.2**	57.8 $\pm$ 0.5	**78.8 $\pm$ 0.3**	38.8 $\pm$ 0.6	**61.7 $\pm$ 0.4**

4.4.4. Analysis of Parameters, FLOPs, and Inference Time

To evaluate the practical efficiency of our BiASTM-CL, we compare it with representative few-shot action recognition methods in terms of parameter count, FLOPs, and inference time in Table 9.

Table 9. Comparison of practical efficiency with representative few-shot action recognition methods under the 5-way 1-shot setting.

Method	Params (M)	FLOPs (G)	Inference Time (ms)
OTAM	25.6	198	112.4
TRX	47.1	205	96.3
STRM	45.5	201	108.6
BiASTM-CL (ours)	51.3	213	103.7

As shown in Table 9, our method requires 51.3 M parameters and 213 G FLOPs. Compared with TRX and STRM, BiASTM-CL introduces only a moderate increase in model complexity. Meanwhile, the reported inference time remains comparable, suggesting that the proposed method maintains practical efficiency.

4.4.5. Visualization Analysis

Under the 5-way 5-shot setting, Figure 7 shows two confusion matrices. Each matrix exhibits a strong diagonal dominance, indicating that the model can still separate classes well under few-shot evaluation. The off-diagonal values are not widely dispersed across many classes; instead, most errors are concentrated in confusions between a small set of visually similar categories. Moreover, the diagonal values are relatively consistent across classes with only small variations, suggesting that the model does not show a strong bias toward any single class. Overall, the confusion matrices follow the accuracy trends reported in the tables and further support the stability and reliability of the model's predictions.

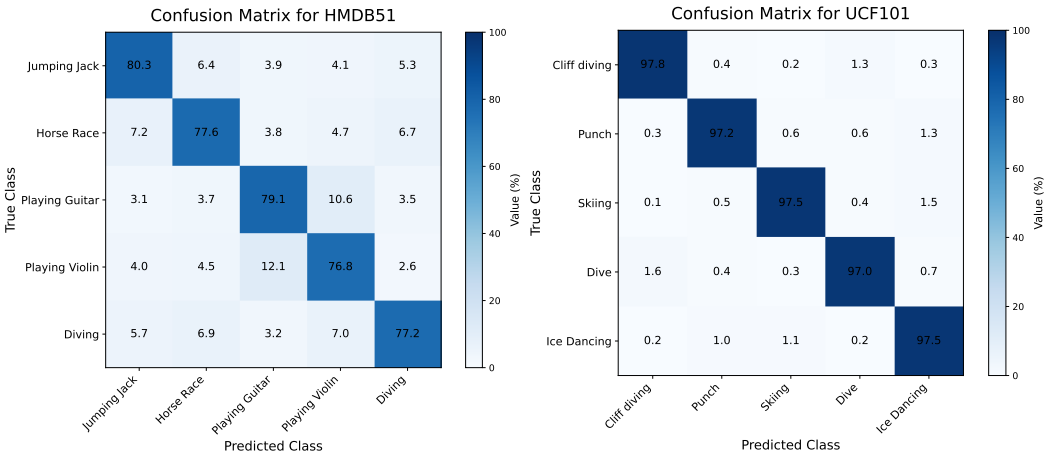


Figure 7. Confusion matrices on HMDB51 and UCF101 under the 5-way 5-shot setting.

To examine whether the model can focus on key frames and fine-grained regions when visually similar actions differ semantically, we visualize attention heatmaps on two representative sequences and compare the results with those of TRX and STRM. Figure 8 shows the comparison for a golf swing sequence.

Figure 9 further presents the attention heatmap comparison for a baseball hitting sequence.

The two sets of heatmaps indicate that BiASTM-CL forms a continuous hotspot aligned with the downswing arc and the impact frame in the golf case, while in the baseball case the attention converges along the horizontal hitting direction to fine-grained regions such as the hand and bat head. Compared with TRX/STRM, our method exhibits less inter-frame jitter, sharper (narrower) peaks, and lower responses to background clutter (e.g., distant scenery or the net). By capturing the semantic difference pattern of "arc vs. horizontal"

while maintaining coordination between fine details and global structure, our approach demonstrates more stable discriminability on this pair of visually similar actions.

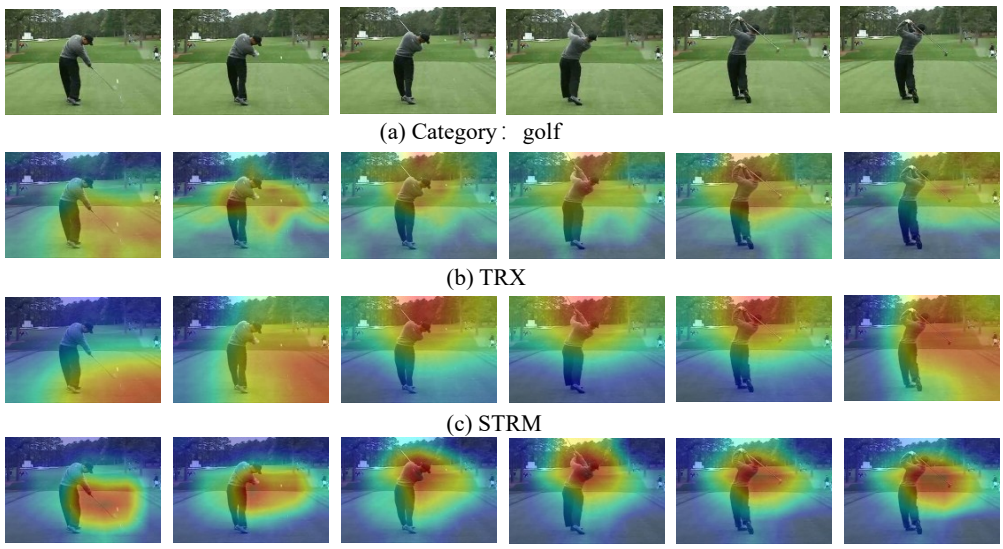


Figure 8. Visual comparison of attention heatmaps for a golf swing. Our method exhibits high responses that continuously shift along the "arm–club head–downswing arc" and align with the impact frame, while maintaining low responses to the distant background; TRX/STRM show peak delays, inter-frame drift, or background attraction.

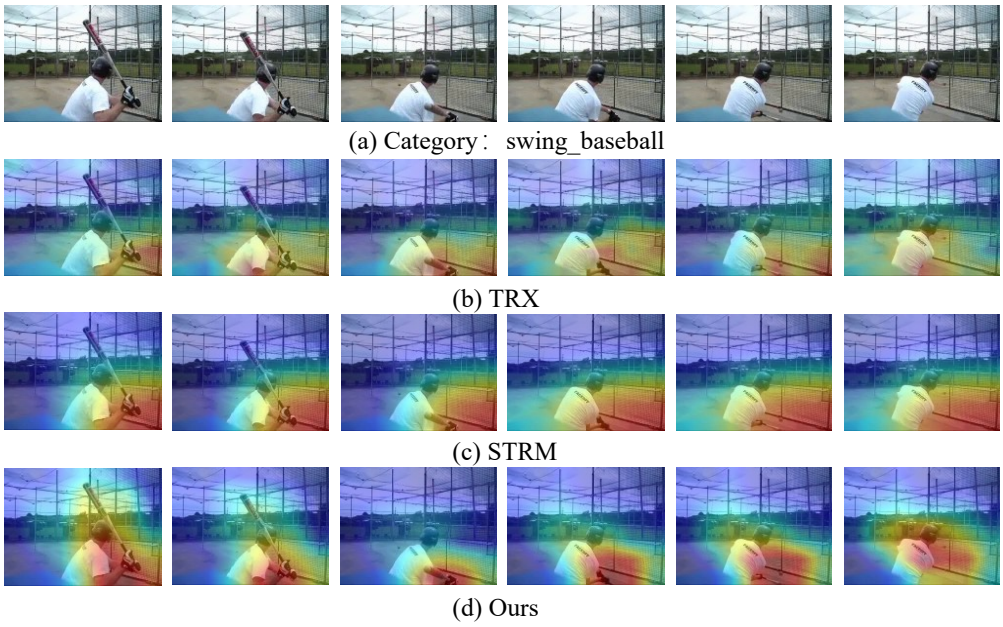


Figure 9. Attention heatmap comparison for a baseball hitting sequence. Our model produces band-like hotspots along the horizontal hitting direction and suppresses interference from highly textured backgrounds (e.g., surrounding netting). In contrast, the two baselines yield more diffuse responses with less emphasis on key regions.

5. Conclusions

We propose a few-shot action recognition framework that integrates bidirectional adaptive spatiotemporal modeling with contrastive learning. The proposed framework is designed to capture local motion variations and global semantic relationships in videos under limited labeled data. It consists of three main components: Attention-Guided Inter-

Frame Differencing (AGID), Local–Global Interactive Channel Attention (LGICA), and Dynamic Distance-Adjusted Contrastive Learning (DACM). AGID improves temporal consistency through inter-frame attention, LGICA enhances representation quality by jointly modeling local and global channel relations, and DACM refines similarity estimation by reducing interference from similar sub-sequences.

Experimental results on UCF101, HMDB51, and SSV2 show that the proposed method achieves the best performance under the 5-shot setting while remaining competitive under the 1-shot setting. Ablation studies further verify the effectiveness of AGID, LGICA, and DACM.

Overall, the proposed method provides an effective and efficient solution for few-shot action recognition, demonstrating that the combination of spatiotemporal feature refinement and contrastive learning is beneficial for this task.

Author Contributions: Conceptualization, Z.Z.; methodology, Z.Z.; software, Z.Z.; validation, Z.Z.; formal analysis, Z.Z.; data curation, Z.Z.; writing—original draft preparation, Z.Z.; writing—review and editing, Z.Z. and J.H.; visualization, Z.Z.; resources, J.H.; supervision, J.H.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Zhejiang Provincial Department of Science and Technology under the Zhejiang Provincial Key R&D Plan (grant numbers: 2023C01023 and 2024C01134).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the following public repositories: UCF101 dataset, available online: https://www.crcv.ucf.edu/research/data-sets/ucf101/ (accessed on 25 April 2025); HMDB51 dataset, available online: https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/ (accessed on 10 June 2025); and Something-Something V2 dataset, available online: https://www.kaggle.com/datasets?search=something-something+v2 (accessed on 7 December 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Verma, K.K.; Singh, B.M.; Dixit, A. A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system. *Int. J. Inf. Technol.* **2022**, *14*, 397–410. [CrossRef]
2. Pang, G.; Shen, C.; Cao, L.; van den Hengel, A.V.D. Deep learning for anomaly detection: A review. *ACM Comput. Surv.* **2021**, *54*, 38. [CrossRef]
3. Ahn, J.; Park, J.; Lee, S.S.; Lee, K.H.; Do, H.; Ko, J. SafeFac: Video-based smart safety monitoring for preventing industrial work accidents. *Expert Syst. Appl.* **2023**, *215*, 119397. [CrossRef]
4. Reddy, S.N.; Kurrey, V.; Nagar, M.; Gupta, G.R. Action recognition based industrial safety violation detection. *arXiv* **2024**, arXiv:2412.05531. [CrossRef]
5. Núñez-Marcos, A.; Arganda-Carreras, I. Transformer-based fall detection in videos. *Eng. Appl. Artif. Intell.* **2024**, *132*, 107937. [CrossRef]
6. Liu, Y.; Boels, M.; Garcia-Peraza-Herrera, L.C.; Vercauteren, T.; Dasgupta, P.; Granados, A.; Ourselin, S. Lovit: Long video transformer for surgical phase recognition. *Med. Image Anal.* **2025**, *99*, 103366. [CrossRef] [PubMed]
7. Santra, S.; Chudasama, V.; Wasnik, P.; Balasubramanian, V.N. Precise Event Spotting in Sports Videos: Solving Long-Range Dependency and Class Imbalance. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 18–24 June 2025; pp. 3163–3172.
8. Lv, Z.; Poiesi, F.; Dong, Q.; Lloret, J.; Song, H. Deep learning for intelligent human–computer interaction. *Appl. Sci.* **2022**, *12*, 11457. [CrossRef]
9. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4489–4497.

10. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NIPS 2014)*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27, pp. 567–575.
11. Wang, L.; Tong, Z.; Ji, B.; Wu, G. Tdn: Temporal difference networks for efficient action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021; pp. 1895–1904.
12. Liu, Y.; Yuan, J.; Tu, Z. Motion-driven visual tempo learning for video-based action recognition. *IEEE Trans. Image Process.* **2022**, *31*, 4104–4116. [CrossRef] [PubMed]
13. Zheng, S.; Chen, S.; Jin, Q. Few-shot action recognition with hierarchical matching and contrastive learning. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 297–313.
14. Wanyan, Y.; Yang, X.; Dong, W.; Xu, C. A Comprehensive Review of Few-Shot Action Recognition. *Int. J. Comput. Vis.* **2025**, *133*, 6832–6859. [CrossRef]
15. Thatipelli, A.; Narayan, S.; Khan, S.; Anwer, R.M.; Khan, F.S.; Ghanem, B. Spatio-temporal relation modeling for few-shot action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 19958–19967.
16. Wang, X.; Zhang, S.; Qing, Z.; Gao, C.; Zhang, Y.; Zhao, D.; Sang, N. Molo: Motion-augmented long-short contrastive learning for few-shot action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 18011–18021.
17. Cai, Q.; Pan, Y.; Yao, T.; Yan, C.; Mei, T. Memory matching networks for one-shot image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4080–4088.
18. Zhang, B.; Li, X.; Feng, S.; Ye, Y.; Ye, R. MetaNODE: Prototype optimization as a neural ODE for few-shot learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 9014–9021.
19. Snell, J.; Swersky, K.; Zemel, R.S. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4077–4087.
20. Li, C.; Zhang, J.; Wu, S.; Jin, X.; Shan, S. Hierarchical compositional representations for few-shot action recognition. *Comput. Vis. Image Underst.* **2024**, *240*, 103911. [CrossRef]
21. Huang, Y.; Yang, L.; Chen, G.; Zhang, H.; Lu, F.; Sato, Y. Matching Compound Prototypes for Few-Shot Action Recognition. *Int. J. Comput. Vis.* **2024**, *132*, 3977–4002. [CrossRef]
22. Perrett, T.; Masullo, A.; Burghardt, T.; Mirmehdi, M.; Damen, D. Temporal-relational crosstransformers for few-shot action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021; pp. 475–484.
23. Ding, Y.; Liu, Y. A novel few-shot action recognition method: Temporal relational CrossTransformers based on image difference pyramid. *IEEE Access* **2022**, *10*, 94536–94544. [CrossRef]
24. Wang, X.; Zhang, S.; Qing, Z.; Tang, M.; Zuo, Z.; Gao, C.; Sang, N. Hybrid relation guided set matching for few-shot action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 19948–19957.
25. Li, S.; Liu, H.; Qian, R.; Li, Y.; See, J.; Fei, M.; Lin, W. Ta2n: Two-stage action alignment network for few-shot action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 1404–1411.
26. Cao, K.; Ji, J.; Cao, Z.; Chang, C.Y.; Niebles, J.C. Few-shot video classification via temporal alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 10618–10627.
27. Lee, S.; Moon, W.; Seong, H.S.; Heo, J.P. Temporal alignment-free video matching for few-shot action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 18–24 June 2025; pp. 5412–5421.
28. Yu, B.; Hou, Y.; Guo, Z.; Gao, Z.; Li, Y. FTAN: Frame-to-frame temporal alignment network with contrastive learning for few-shot action recognition. *Image Vis. Comput.* **2024**, *149*, 105159. [CrossRef]
29. Tian, J.; Gu, J.; Pu, Y.; Zhao, Z. Skeleton-Based Few-Shot Action Recognition via Fine-Grained Information Capture and Adaptive Metric Aggregation. *IEEE Trans. Instrum. Meas.* **2024**, *74*, 5002812. [CrossRef]
30. Guo, F.; Zhu, L.; Wang, Y.; Sun, J. Task-specific alignment and multiple-level transformer for few-shot action recognition. *Neurocomputing* **2024**, *598*, 128044. [CrossRef]
31. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 20–36.
32. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human action classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402. [CrossRef]

33. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the 2011 International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.
34. Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. The "something something" video database for learning and evaluating visual common sense. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5842–5850.
35. Xing, J.; Wang, M.; Liu, Y.; Mu, B. Revisiting the spatial and temporal modeling for few-shot action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2023; Volume 37, pp. 3001–3009.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.