

MoVie: Broaden Your Views with Human Motion for Action Detection

Di Yang¹ Mahmoud Ali² Xuanlong Yu³ Xi Shen³ Quan Kong⁴
Gianpiero Francesca⁵ François Brémond²

¹Suzhou Institute for Advanced Research, University of Science and Technology of China, China

²Inria Center at Université Côte d’Azur, France

³Intellindust AI Lab, China ⁴Woven by Toyota, Japan ⁵Toyota Motor Europe, Belgium

Abstract

Human action detection in videos requires both semantic recognition and accurate modeling of motion. While recent video foundation models have advanced visual semantics, they still struggle to capture complex and compositional actions due to the limited representation ability of motion. Human skeleton sequences, which explicitly describe the body structure and movement, provide valuable physical and geometric motions that complement RGB videos. However, combining video and skeleton modalities faces two key challenges: (i) label-driven skeleton features are too coarse to describe fine-grained motion, and (ii) skeleton motion and RGB video lie in heterogeneous feature spaces, so current fusion strategies often cause feature interference. To address these, we propose MoVie¹, a unified Motion-Video processing framework that uses structured human motion as a bridge between the two signals. We first propose a Structural Motion Projection module that decomposes motion into primitive components using a learnable motion dictionary, to produce fine-grained descriptors. Then, we design a Motion-guided Feature Regularization mechanism that aligns visual features with motion through an orthogonality-based transformation, so that fine-grained motion cues can guide visual representations without collapsing semantic diversity. Extensive evaluations on Toyota Smarthome Untrimmed, Charades, Multi-THUMOS and PKU-MMD datasets demonstrate that MoVie significantly improves state-of-the-art action detection performance.

1. Introduction

Human action detection is a crucial task in video understanding [2, 5, 19, 20, 23, 24, 48, 56]. The goal is to identify and localize actions in untrimmed videos by assigning multi-class and frame-wise labels. Recent state-of-the-art methods [11, 12, 14, 44, 63] build temporal modeling on visual features extracted from powerful video foundation

models that learn from videos [21, 30, 46, 49, 58] and semantics [28, 33, 34, 50]. However, these models still struggle in fine-grained and multi-label action detection in complex scenes. The main reason is that they ignore explicit *motion* that provides the physical and geometric essence of human actions. For instance, actions performed under varying subjects, viewpoints, or lighting conditions may look indistinguishable in RGB space but differ in their sequential motion dynamics. Without modeling such dynamics, these models can describe what is visible in a scene, but not how the actions physically unfold over time.

To explicitly capture human motion, *skeleton data*, represented as 2D or 3D human keypoints, provide a valuable modality [18, 38, 43, 53, 56, 59]. It describes the physical structure and movement of the human body directly. This information can help visual models perceive motion patterns more accurately. However, most existing action detection methods [11, 12, 63] have not fully leveraged skeleton motion as complementary information. In practice, simply introducing motion as an additional modality or fusing it with RGB features brings limited improvement [9, 62]. This suggests that the challenge lies not in adding another data source, but in understanding how motion and visual representations should be structurally related for effective action modeling.

The first problem is that the motion representation is often too coarse to be combined with the visual features. Most existing skeleton-based encoders [6, 56] are trained with global action labels. They learn which category an action belongs to, but they do not learn the native structure of motion itself. The resulting motion features mix different physical patterns and fail to capture the fine structure that corresponds to real human dynamics. This approximation makes it difficult for motion information to effectively guide visual models. The second problem is that the motion representation itself lacks structure. In the real world, actions are complex. They are formed by combinations of smaller motion units, such as raising an arm, bending, or stepping forward. These smaller motions, called motion primitives, can overlap and recombine into complex behaviors. With-

¹<https://walker1126.github.io/MoVie-project>

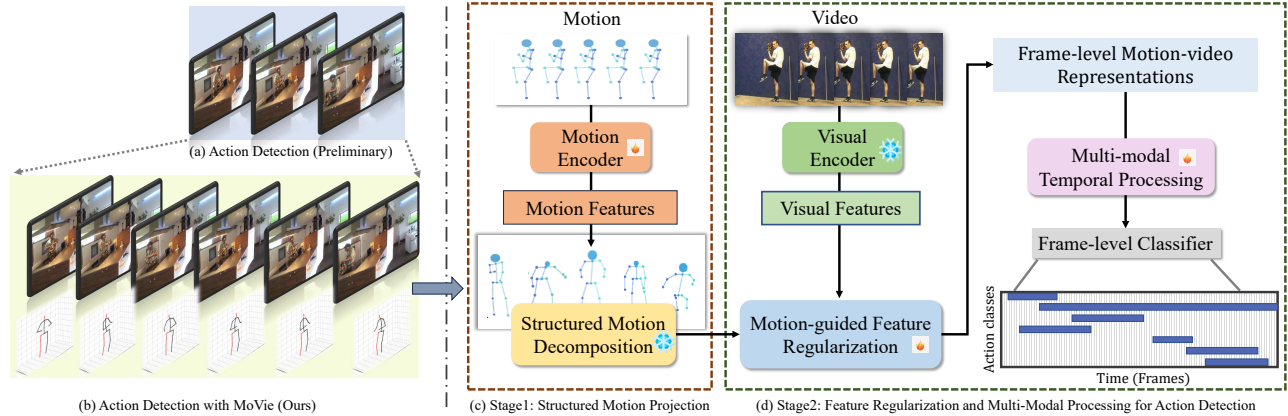


Figure 1. **Overview of the MoViE framework.** (a) Conventional RGB-based methods struggle to capture complex and compositional motions. (b) MoViE introduces motion as a structured physical prior to complement visual representation. (c) Structured motion primitives are learned from skeleton sequences through a motion decomposition model. (d) The Motion-guided Feature Regularization (MGFR) aligns these primitives to inject geometric and physical cues into the visual space, and the learned features are sent to a history-aware Multi-modal Temporal Processing for more accurate action detection.

out decomposing actions into such primitives, models cannot match the fine-grained nature of motion and generalize to unseen combinations.

To solve these problems, we propose a unified motion-video framework, namely MoViE, to treat skeleton motion not as an auxiliary input but as a structural bridge that connects physical movement with visual representation. Our goal is to make the visual model understand motion in a structured and fine-grained manner. We design a two-stage framework to achieve this.

As shown in Fig. 1, in the first stage, we build a Structural Motion Projection (SMP) module. This module learns to describe motion in terms of its basic components. Inspired by the idea of motion decomposition [57, 59], we project skeleton features onto a set of learnable motion bases, called a motion dictionary. Each basis represents a basic motion pattern, and the model learns how much each pattern is activated in a given sequence. These activation magnitudes reflect the strength of each primitive motion. We use them to form a new motion representation that focuses on how a person moves, not what coarse action label they belong to. This stage builds a structured connection between motion and visual space.

In the second stage, we introduce a Motion-guided Feature Regularization (MGFR) module. This module uses the structured motion representation to adjust how visual features change with different actions. Instead of directly merging motion and visual features, we let motion act as a regulator to align the spatial representation of visual features with the structured fine-grained motion signals derived from the learned viewpoint-invariant motion dictionary. After regularization, the combined motion-visual features are sent to a history-aware Temporal Modeling, which processes previously stored visual features alongside the

learned motion-visual features. Finally, a multi-label classifier assigns action categories to each frame, for frame-wise action detection.

We validate the effectiveness of our approach through extensive experiments on datasets including real-world challenges *e.g.*, viewpoints/subjects variants, action overlapping, and background clutter: Toyota Smarthome Untrimmed [13], Charades [42], Multi-THUMOS [60], and PKU-MMD [7]. Our results demonstrate that the motion-improved representations significantly improve the performance of action detection, which surpasses state-of-the-art methods.

In summary, the contributions of this paper are as follows: (i) We introduce MoViE, a novel framework to learn structured primitive motion representation and dynamic feature alignment with visual models. (ii) We propose a Structural Motion Projection module that extracts primitive magnitudes from a motion dictionary, transforming skeleton motion into a structured, physically meaningful representation. (iii) We design a Motion-Guided Feature Regularization mechanism that dynamically aligns visual representations with fine-grained motion via orthogonal basis re-parameterization. (iv) Extensive experiments are conducted on multiple challenging real-world datasets, focusing on frame-wise multi-label action detection. We show that MoViE effectively establishes motion as a structural bridge rather than a secondary modality, and enhances dynamic consistency and interpretability in action detection.

2. Related work

Human Temporal Action Detection: Human temporal action detection focuses on classifying activities frame by frame in untrimmed videos, where multiple actions can oc-

cur simultaneously. The main challenge is how to model long-term relationships between activities at different time points and how to handle complex activities. Most current methods use untrimmed RGB videos. Since these videos often contain thousands of frames, training a single deep neural network on such videos is very expensive. Previous works have proposed a two-step approach to address this issue. In the first step, a pre-trained feature extractor (e.g., I3D [5]) is applied to short video segments to extract visual features. In the second step, action detection is treated as a sequence-to-sequence (seq2seq) task, translating visual features into per-frame action labels. Temporal Convolution Networks (TCNs) [10, 11, 25, 26, 61] and Transformers [12, 14] are commonly used in the second step because they capture long-term dependencies. Recently, some methods have focused on improving temporal modeling in the second stage. For example, MLAD [45] uses attention mechanisms to model actions that happen simultaneously and at different times. DualDETR [63] introduces a dual-level framework to better explore the potential of a query-based method.

Several recent methods [9, 13] have started to use skeleton motion in this task to benefit from multi-modal information. In these methods, a pre-trained Graph Convolutional Network (GCN), such as AGCN [38], is used as a visual encoder to extract skeleton features. However, unlike pre-trained I3D, which works well across different datasets, pre-trained AGCN struggles to provide high-quality features because it was trained on the NTU-RGB+D dataset [37], which is designed for controlled environments. We found that the performance drops significantly when applying the pre-trained AGCN model to more complex real-world datasets, such as TSU [13] and Charades [42].

MoVie differs from previous methods, which focus on features from visual and semantic modalities. We propose a motion-regularized module that encodes visual features within a skeleton encoder without greatly increasing the model size, and view motion as a structural regulator that organizes visual dynamics in a physically consistent way. Our design captures the intrinsic geometry and dynamics of human movement, and leverages it to regularize temporal feature evolution in RGB models.

Multi-modal Video Representation: Multi-modal video representation learning aims to improve video representation ability by combining features from different modalities. Recently, many methods have used language features [33] for video understanding [27, 29, 39, 40, 47, 52], video captioning [54], and visual question answering [3, 41]. However, while semantic information is generally helpful, it may not be as effective for complex human motion-oriented tasks [1, 14]. To improve human action representation ability, current methods in motion-visual models focus on creating frameworks that can effectively capture and understand the relationships between visual and motion data in

video sequences [8]. State-of-the-art computer vision techniques with advanced motion modeling [13, 17] have used attention mechanisms [16] or distillation techniques [17] to combine both RGB features and skeleton motion features. However, motion-augmented methods [13, 15, 17] using skeleton data are mainly designed for short videos, and such methods are sensitive to the feature quality from multi-modal data. The ability to handle real-world actions over longer time periods is still insufficient.

Unlike previous methods that combine global visual and skeleton features, MoVie introduces a different perspective. It constructs a Structured Motion Projection to decompose motion into interpretable primitives, capturing the underlying physical and geometric structure of actions. This structured motion is then used in a Motion-Regulated Feature Regularization process to guide RGB features toward dynamically consistent representations. This design allows motion to act as a structural bridge that connects physical movement with visual perception, to offer a unified and physically grounded understanding of human actions.

3. Proposed Approach

In this section, we present the full architecture of the proposed MoVie framework.

Overview Architecture: Given a video segment \mathbf{v} sampled from an untrimmed video, we first extract visual features \mathbf{F}_v using a frozen visual encoder E_V , such as I3D [5] or ViCLIP [50]. Simultaneously, we obtain 2D/3D human skeleton sequences \mathbf{m} from a pose estimator [4, 35, 55]. In the *Structured Motion Projection (SMP)* module, each skeleton sequence is processed by a pretrained ViA motion decomposition network [59] to extract and project structured motion features \mathbf{F} . The motion and visual representations are then aligned through a *Motion-Guided Feature Regularization (MGFR)* module, which maps structured motion primitives into the visual feature space to correct and stabilize visual dynamics. The resulting motion-regularized visual features are then processed through multi-person pooling and cross-modal temporal modeling to achieve frame-level representations for action detection.

3.1. Visual Feature Extraction

For action detection, the input video is an untrimmed video that may span a long duration [13]. As a compromise, similar to previous action detection models [57], we denote a visual encoder E_V , which could be I3D [5], or a CLIP-based video foundation model [50]. For each segment \mathbf{v} , we extract concatenated per-frame features \mathbf{F}_v as in Eq. 1:

$$\mathbf{F}_v = E_V(\mathbf{v}). \quad (1)$$

where $\mathbf{F}_v \in \mathbb{R}^{C_v \times T}$ and T denotes the number of frames in the processed window to extract the video segments.

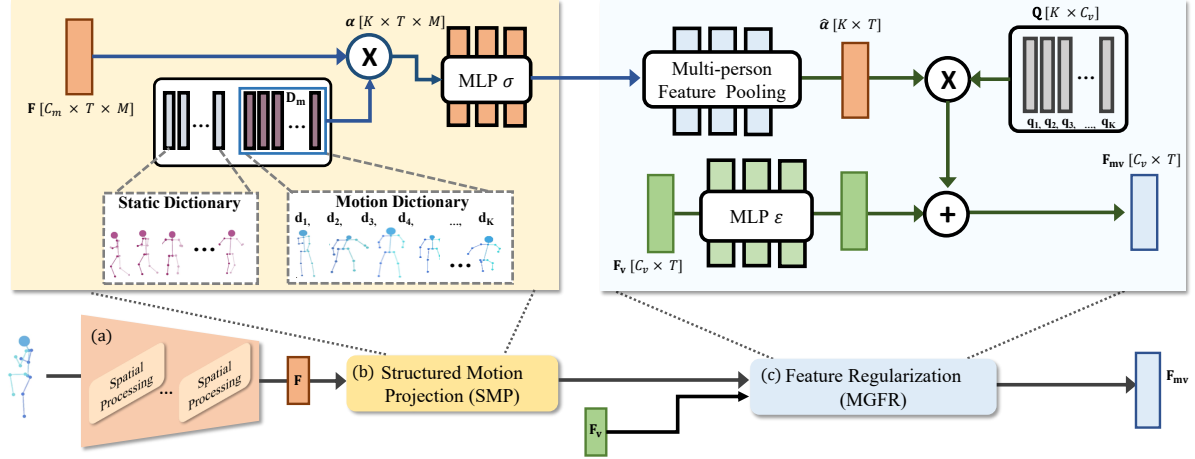


Figure 2. **Overview of the Regularized Motion-Video Feature Learning.** Given an input skeleton sequence, we obtain the features $\mathbf{F} \in \mathbb{R}^{C_m \times T \times M}$ via (a) a spatio-temporal motion encoder. The (b) Structured Motion Projection (SMP) module projects \mathbf{F} onto a pretrained motion dictionary in the prior stage, $\mathbf{D}_m \in \mathbb{R}^{K \times C_m}$, and computes activation magnitudes $\alpha \in \mathbb{R}^{K \times T \times M}$. These structured coefficients α represent the primitive-level motion strength that is invariant to viewpoint and appearance. Subsequently, the (c) Motion-Guided Feature Regularization (MGFR) aligns the projected structured motion descriptors $\hat{\alpha}$ with the RGB visual features $\mathbf{F}_v \in \mathbb{R}^{C_v \times T}$ by an orthogonal projection $\mathbf{Q} \in \mathbb{R}^{K \times C_v}$. The output \mathbf{F}_{mv} are used for frame-wise action detection by Temporal Modeling.

As shown in Fig. 2, such visual features \mathbf{F}_v will be fed into the proposed Motion-guided Feature Regularization module and at the same time, stored in a fixed memory bank as a history to be reused and updated in the next training iteration.

3.2. Structural Motion Projection

Skeleton motion carries explicit geometric and kinematic information of human actions, this section presents the motion primitive feature extraction and projection approach.

Motion Feature Extraction: The skeleton modality can be obtained from various sources such as wearable sensors, RGB-D cameras, or human pose estimation models. In real-world videos, we extract skeletons using a pose estimation algorithm [55], which achieves high quality even under challenging conditions such as occlusion or clutter. More implementation details and skeleton quality analysis are provided in the Appendix.

Given a skeleton sequence, we process it as a spatio-temporal tensor $\mathbf{m} \in \mathbb{R}^{C_{in} \times T \times M \times J}$, where T is the number of frames in the temporal window, M is the number of detected persons, J is the number of joints, and C_{in} denotes the coordinate channels (2D or 3D). To effectively capture motion features \mathbf{F} through time and space, we employ a stack of spatio-temporal layers following [56]. Each spatial layer uses a multi-head attention mechanism to model the dependencies among body joints, while the temporal layers employ temporal convolutional networks (TCNs) to capture the motion dynamics across frames. After several stages of spatial-temporal processing, we obtain per-frame,

per-person motion features:

$$\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_T], \quad \mathbf{f}_t = \{\mathbf{f}_t^{(1)}, \dots, \mathbf{f}_t^{(M)}\}, \quad (2)$$

where $\mathbf{f}_t^{(p)} \in \mathbb{R}^{C_m}$ denotes the motion embedding of person p at frame t .

Structured Motion Decomposition and Projection: To obtain a robust and physically interpretable motion representation, we leverage a pretrained motion decomposition encoder [59], which learns two orthogonal subspaces: a *motion dictionary* $\mathbf{D}_m \in \mathbb{R}^{K \times C_m}$ and a static one capturing appearance-related variations. We only use the motion dictionary \mathbf{D}_m , where each basis vector represents a primitive movement direction that is invariant to viewpoint and body shape. The dictionary is pretrained through cross-view motion reconstruction, ensuring that each primitive captures intrinsic kinematic properties of human motion.

Given a skeleton feature tensor $\mathbf{F} \in \mathbb{R}^{C_m \times T \times M}$, where T and M denote the number of frames and detected persons, respectively, we project the motion features onto the learned motion dictionary. For each frame t and person p , we compute the primitive-level activation magnitudes as:

$$\alpha = \|\mathbf{D}_m \mathbf{F}\|_2, \quad \alpha \in \mathbb{R}^{K \times T \times M}, \quad (3)$$

where each element α_k denotes the activation strength (amplitude) of the k -th motion primitive at frame t for person p . This representation describes structured motion magnitudes that are disentangled from static appearance and invariant to camera view, since each motion primitive encodes only the geometric and kinematic dynamics.

To further adapt motion features to the visual feature space, we employ a light-weight MLP projection $\sigma(\cdot)$:

$$\tilde{\alpha} = \sigma(\alpha), \quad (4)$$

where $\tilde{\alpha} \in \mathbb{R}^{K \times T \times M}$ serves as the refined structured motion descriptor.

Multi-person Pooling: In real-world videos, multiple people may appear simultaneously. We aggregate motion features across detected persons to form a unified per-frame descriptor. The per-person motion coefficients $\tilde{\alpha}$ are first projected by $\sigma(\cdot)$ to stabilize noise, and then merged through a pooling operator \mathcal{G}_p :

$$\hat{\alpha} = \mathcal{G}_p(\sigma(\alpha)) \in \mathbb{R}^{K \times T}. \quad (5)$$

\mathcal{G}_p performs a linear projection followed by max or mean pooling over the person dimension. We select the top- M skeletons by pose confidence and pad zeros when fewer are available. The resulting $\hat{\alpha}$ provides a stable structured motion descriptor for the subsequent *Motion-Guided Feature Regularization (MGFR)* module.

3.3. Motion-Guided Feature Regularization

The structured motion features $\hat{\alpha}$ provide geometric and physical priors about human dynamics, while the visual features $\mathbf{F}_v \in \mathbb{R}^{C_v \times T}$ encode rich appearance and semantic context. However, these two feature spaces are inherently heterogeneous: motion primitives describe directional movement magnitudes, whereas visual embeddings represent high-level semantic cues. To establish a structured correspondence, we propose the *Motion-Guided Feature Regularization (MGFR)* module, which aligns motion and vision through two consecutive projections on orthogonal bases.

Orthogonal Projections: The first projection in the Motion Primitive Projection decomposes motion embeddings into a compact set of basis movements using a fixed motion dictionary \mathbf{D}_m (Sec. 3.2), yielding the primitive activation α . This operation can be seen as projecting raw motion onto an interpretable low-dimensional manifold that captures geometry-invariant motion primitives.

Here, we propose that the second projection aligns these primitives with the visual feature space. We introduce another learnable orthogonal transformation $\mathbf{Q} \in \mathbb{R}^{K \times C_v}$ that defines a primitive-aligned coordinate system that allows motion signals to modulate visual features along disentangled motion directions. To ensure stable mapping, we first project both modalities through shallow MLPs, $\sigma(\cdot)$ for motion and $\epsilon(\cdot)$ for vision, to normalize their scales and nonlinearities. The motion-regularized visual features are then obtained as:

$$\mathbf{F}_{mv} = \epsilon(\mathbf{F}_v) + \lambda(\mathbf{Q}^\top \hat{\alpha}), \quad (6)$$

where λ controls the modulation strength. Here, \mathbf{Q} is constrained to be orthogonal:

$$\langle \mathbf{q}_i, \mathbf{q}_j \rangle = \begin{cases} 0, & i \neq j, \\ 1, & i = j, \end{cases} \quad (7)$$

and is re-orthogonalized at each iteration using the Gram-Schmidt process following [58]. Intuitively, Eq. 6 allows visual features to be modulated along physically meaningful motion directions.

Consistency Regularization: To further stabilize the alignment, we introduce a temporal consistency loss that encourages motion-induced changes to match natural variations in the visual features:

$$\mathcal{L}_{align} = \frac{1}{T} \sum_{t=1}^T \left\| \mathbf{Q}^\top \hat{\alpha}_t - (\epsilon(\mathbf{F}_{v,t}) - \mathbf{F}_{mv}^{mean}) \right\|_2^2, \quad (8)$$

where \mathbf{F}_{mv}^{mean} denotes the temporal mean of visual features. This term enforces that the temporal evolution implied by motion primitives aligns with the observed changes in appearance and semantics.

After MGFR, the motion-regularized feature sequence $\mathbf{F}_{mv} \in \mathbb{R}^{C_v \times T}$ is passed to the multi-person pooling and cross-modal temporal modules. Compared to concatenation or late fusion, MGFR injects motion as a structural regularizer in the visual channel space. This design keeps the rich semantic content of visual features while adding geometric and physical consistency.

3.4. Temporal Processing and Framework Training

We feed the motion-regularized per-frame features \mathbf{F}_{mv} to a temporal encoder TM that models both short- and long-range dependencies.

Temporal Modeling: To model temporal dependencies across frames, we follow MS-TCT [12], where TM alternates Transformer and TCN layers to jointly capture global and local dynamics. During training, a sliding window of size W_s is used to process long videos efficiently. We further concatenate the motion-guided visual features \mathbf{F}_{mv} with the cached history feature $\mathbf{F}_h \in \mathbb{R}^{C_v \times T}$ along the channel dimension:

$$\mathbf{F}'_{mv} = \text{TM}(\text{concat}[\mathbf{F}_{mv}, \mathbf{F}_h]). \quad (9)$$

This produces temporally consistent representations for frame-wise action detection, and online inference.

Training Strategy: The motion dictionary is pretrained independently for cross-view motion reconstruction and kept frozen in our training. We train the remaining components end-to-end following $\mathcal{L} = \mathcal{L}_{det} + \lambda_{align} \mathcal{L}_{align}$. As we address frame-wise action detection for videos with dense action occurrences by treating each frame feature, \mathbf{F}'_{mv} , as inputs to a multi-label classification task. A per-frame classifier, stacked on top of \mathbf{F}'_{mv} , generates action prediction

Methods	Modality	Feature	Toyota Smarthome	Untrimmed	Charades	MultiTHUMOS
			CS(%)	CV(%)	mAP(%)	mAP(%)
R-C3D [51]	Visual	C3D	8.7	-	17.6	-
Super-event [31]	Visual	I3D	17.2	-	18.6	36.4
TGM [32]	Visual	I3D	26.7	-	20.6	37.2
SD-TCN [13]	Visual	I3D	29.2	18.3	21.6	-
PDAN [11]	Visual	I3D	32.7	-	23.7	40.2
Coarse-Fine [25]	Visual	X3D	-	-	25.1	-
MLAD [45]	Visual	I3D	-	-	18.4	42.2
MS-TCT [12]	Visual	I3D	33.7	-	25.4	43.1
DualDETR [63]	Visual	I3D	34.8	-	23.2	45.5
TTM [36]	Visual	ViViT-L	-	-	28.8	-
Bi-LSTM [22]	Motion	LSTM	17.0	14.8	8.2	-
TGM [32]	Motion	TGM	26.7	13.4	9.0	-
SD-TCN [13]	Motion	AGCN	26.2	22.4	9.8	-
LAC [57]	Motion	UNIK	36.8	23.1	25.6	23.4
Augmented-RGB [9]	Flow&Motion&Visual	I3D	32.8	24.6	-	44.6
MoVie (Ours)	Motion&Visual	I3D	49.6	28.6	29.2	46.8
PDAN [11]	Visual	ViCLIP	21.5	13.4	16.1	33.5
MS-TCT [12]	Visual	ViCLIP	35.8	-	16.4	39.2
AAN [14]	Text&Visual	CLIP	41.3	-	32.0	-
MMFF [62]	Motion&Visual	ViCLIP	41.6	25.7	29.2	46.3
MoVie (Ours)	Motion&Visual	ViCLIP	50.1	30.1	33.5	48.3

Table 1. Frame-level mAP on TSU, Charades and Multi-THUMOS for comparison with SoTA action detection methods. Modalities used by the approaches are shown for reference.

scores denoted by $P \in \mathbb{R}^{Cls \times T}$. We optimize the framework with a Binary Cross Entropy (BCE) loss:

$$\mathcal{L}_{det} = -\frac{1}{T} \sum_{t=1}^T \sum_{c=1}^{Cls} [y_{t,c} \log P_{t,c} + (1 - y_{t,c}) \log(1 - P_{t,c})], \quad (10)$$

where T is the number of frames, Cls is the number of action classes, and y represents the ground truth. This loss function serves as the primary criterion for frame-level action detection.

4. Experiments and Analysis

This section presents our experimental settings and results. We evaluate MoVie on real-world multi-label videos with frame-level annotations. We first describe the datasets and protocols, then provide comparisons with state-of-the-art approaches, ablation studies, and further analysis. Implementation details, computation cost, and additional analyses are provided in the Supplementary Material (Appendix).

4.1. Datasets and Evaluation Metrics

Toyota Smarthome Untrimmed (TSU) [13] contains long continuous videos with dense frame-wise labels. Up to five actions may appear in the same frame. We follow the official Cross-Subject (CS) and Cross-View (CV) splits and report frame-level mAP. The dataset provides 2D skeletons, and we keep the same input format as prior work.

Charades [42] features indoor daily activities with strong ambiguity between actions. The dataset does not include skeleton data. We extract 2D poses using the method

in [55]. We adopt the standard frame-level mAP as the evaluation metric.

Multi-THUMOS [60] extends THUMOS14 with dense annotations over 65 classes. Many actions are short and appear in rapid succession. We extract 2D poses using [55]. We follow the setting of [12] and report frame-level mAP.

PKU-MMD [7] To examine performance in a different temporal regime, we further evaluate MoVie on PKU-MMD part I. The dataset contains long activities divided into event segments. We follow the Cross-Subject (CS) protocol and report event-level mAP at IoU 0.1.

4.2. Comparison with State-of-the-Art

We evaluate MoVie on three challenging benchmarks TSU, Charades, and Multi-THUMOS, and compare it with recent state-of-the-art (SoTA) approaches under different modalities. Results are summarized in Table 1.

Overall Performance: MoVie consistently outperforms all baselines across datasets and feature backbones. With I3D features, MoVie outperforms the previous SoTA [12] by +15.9% on TSU-CS and +3.7% mAP on Multi-THUMOS. On Charades, our model shows higher accuracy than SoTA [36] that based on ViViT [2] features.

Effectiveness of Motion Representation: Compared to motion-only models such as LAC [57] and SD-TCN [13], MoVie achieves significantly higher accuracy (*e.g.*, +12.8% on TSU-CS). This confirms that structured motion is most effective when used to guide visual representations rather than being used alone. The improvements on the Charades

Model Variants	TSU CS (%)	Charades (%)
Fusion Strategy and MGFR Effect		
Baseline (Visual only)	35.8	16.4
Late Fusion	37.1	20.8
Concatenation [F_v, F]	41.2	29.3
MGFR only (w/ F)	44.1	29.6
SMP and MGFR Synergy		
SMP+MGFR w/ $K=64$	41.4	30.4
SMP+MGFR w/ $K=128$	50.1	33.5
SMP+MGFR w/ $K=256$	49.6	33.1
Orthogonality Analysis		
SMP+MGFR w/o Orth.	47.3	31.1
SMP+MGFR w/ Orth.	50.1	33.5

Table 2. Ablation study on MGFR, SMP, and orthogonality on TSU and Charades datasets.

and Multi-THUMOS datasets without native skeletons also validate the robustness of our method when skeletons are estimated automatically, and shows that motion provides generalizable cues even under noisy pose extraction.

Comparison with Other Modalities: Compared with visual-flow and visual-text fusion models [9, 14], MoVie achieves higher mAP despite using fewer modalities. This highlights that physical and geometric priors from structured motion complement visual semantics more effectively than language-based cues for fine-grained temporal understanding. Unlike prior visual-motion fusion approaches [62] that rely on attention-based concatenation, our MGFR aligns motion primitives with RGB features in an orthogonal subspace, providing a more stable and interpretable enhancement to visual representations.

In summary, MoVie establishes new SoTA results across all benchmarks, demonstrating that introducing structured motion as a geometric prior effectively improves long-term temporal reasoning and fine-grained action discrimination.

4.3. Ablation Study

To assess the effectiveness of each component in our framework, we perform ablation studies on the TSU (CS) and Charades datasets. For each ablation study on an individual module, we use the full model and keep other modules at their default optimal settings (see Appendix for details on parameter settings).

Fusion Strategy and Effect of MGFR: We first compare different strategies for integrating motion and visual information (see Tab. 2 (top)). Simple late fusion or feature concatenation yields only moderate gains over the visual baseline, showing that naive combination cannot fully exploit the motion cues. In contrast, our Motion-Guided Feature Regularization (MGFR) achieves consistent improvements (+8.3% on TSU-CS and +13.2% on Charades), demonstrating that learning an orthogonal projection from motion primitives to the visual feature space enables motion to act as a structured regulator rather than a redundant signal. This

Model Variants	TSU CS (%)	Charades (%)
Interaction Mechanism		
Single person	48.3	32.3
Average Pooling	49.6	33.1
MLP Pooling	50.1	33.5
History Features		
w/o history	49.0	32.6
w/ attention	49.9	32.8
w/ concatenation	50.1	33.5

Table 3. Ablation on interaction pooling and history features on TSU and Charades datasets.

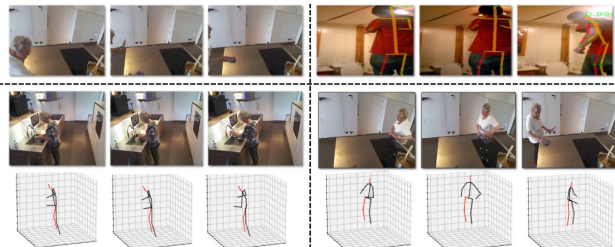


Figure 3. Examples of video with challenging light conditions (top right) and confused background (bottom right), video without accurate skeleton data in large occlusions (top left), video with the simple scenario (bottom left). MoVie can improve the performance compared to single-modal approaches in all these cases.

alignment allows motion cues to correct and stabilize temporal dynamics in visual representations.

Synergy between SMP and MGFR: We further study the joint effect of Structured Motion Projection (SMP) and MGFR. Using MGFR alone provides noticeable improvements, but the performance is significantly enhanced when the motion is first decomposed into primitives through SMP. Label-driven global motion features fail to capture detailed dynamics, while the primitive-level representation provides fine-grained physical bases that make the alignment in MGFR more effective. As shown in Tab. 2 (middle), SMP+MGFR with $K = 128$ achieves the best results, suggesting that decomposed motion offers sufficient expressiveness without introducing redundant bases. This confirms that SMP provides structured, disentangled motion signals, and MGFR aligns them with RGB semantics in a geometrically consistent space.

Effect of Orthogonality: We further test the importance of the orthogonal constraint in MGFR in Tab. 2 (bottom). Replacing the orthogonal projection with a dense linear layer reduces performance (-2.8% on TSU-CS). Without orthogonality, the model easily overfits and mixes correlated visual channels. The orthogonal constraint keeps the projection space more stable and interpretable. Each motion primitive can then adjust visual features along independent directions, which improves consistency and generalization.

Activity	Gain from MoVie
Get up	+46.9
Stir the pot	+32.8
Sit down	+31.1
Watch TV	+21.5
Clean dishes	+19.5
Mean Accuracy	+9.7
Use tablet	-1.4
Telephone	-1.8
Drink from cup	-3.4
Drink from bottle	-4.1
Stir coffee	-5.3

Table 4. Activities that benefit on frame-level mAP (%) from MoVie on TSU-CS, compared with Visual-only model [12].

Interaction Mechanism and History Features: We analyze the role of interaction pooling and history features. Tab. 3 shows that average pooling gives stable results, while MLP pooling performs best by learning adaptive relations between people. Adding history context brings small gains, and simple concatenation works better than attention. These results show that interaction and temporal history offer useful but minor improvements compared with the main motion-visual alignment.

4.4. Further Study

This section presents further quantitative and qualitative evaluations to analyze the effectiveness of MoVie.

Per-class Comparisons and Analysis: We further analyze per-class improvements on TSU-CS compared with the single-modal baselines. From Fig. 3, we find that the videos with challenging light conditions can benefit largely from MoVie compared to models using RGB data only [12]. However, high-quality skeleton data is sometimes not available in real-world with large occlusions, in these cases, MoVie performs significantly better than models using motion data only [57].

Furthermore, results in Tab. 4 show that MoVie achieves large gains on motion-intensive activities such as “Get up” (+46.9%), “Stir the pot” (+32.8%), and “Sit down” (+31.1%). These actions contain clear and repetitive body dynamics that can be effectively captured by structured motion primitives. Each primitive represents a fixed basis motion with a clear physical meaning (Fig. 4), while the activation coefficient α indicates how strongly that primitive is used in a given video. For the action “Get up”, we compute the activation change $|\Delta\alpha|$ for each primitive over the action segment and visualize the primitives with the largest and smallest variations. Primitives with the largest $|\Delta\alpha|$ correspond to meaningful body movements, such as torso bending (α_8) and leg extension (α_{15}), whereas primitives with minimal variation (α_{83}, α_{64}) mainly capture minor arm motions. This result shows that MoVie selectively

Method	Modality	PKU	TSU
Two-stream [5]	Flow&Visual	83.4	-
Pose-RGB [9]	Motion&Visual	84.7	-
Augmented-RGB [9]	Flow&Motion&Visual	86.3	15.1
MMFF [62]	Motion&Visual	79.6	18.3
MoVie	Motion&Visual	92.8	25.6

Table 5. Event-level detection performance on TSU-CS and PKU-MMD (PKU)-CS. MoVie improves multi-modal baselines.

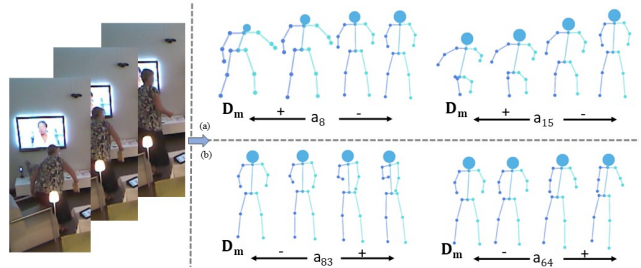


Figure 4. Example of the SMP (“Get up”). Primitives in (a) that correspond to torso bending and leg extension are strongly activated. Primitives in (b) with almost no activation change mostly correspond to arm rotation or minor upper-body motion.

amplifies motion primitives relevant to the action while suppressing less informative movements, leading to improved detection accuracy.

On the other hand, fine-grained actions such as “Drink from bottle” or “Stir coffee” show minor drops, mainly due to weak motion cues or unreliable skeleton estimation under occlusion. These cases suggest that combining hand-object interactions and modeling motion uncertainty could further improve subtle action recognition.

Event-level Evaluation: We further examine MoVie in an event-level setting on TSU and PKU-MMD following cross-subject protocols. This task groups multiple short actions into a long semantic event, which creates longer temporal dependencies. Tab. 5 shows that MoVie improves baselines with different modalities. It suggests that structured motion provides stable complementarities across different temporal scales.

5. Conclusions

We presented MoVie, a motion-augmented framework for real-world human action detection. MoVie learns structured skeleton motion and introduces a simple but effective mechanism to align and fuse primitive-centric motion with visual features. This design allows the model to build more stable representations. Experiments on several challenging datasets show that motion provides clear complementary information to RGB, and that our motion-vision fusion improves frame-level action detection accuracy. MoVie gives a practical direction for building stronger multi-modal action detectors in complex scenarios.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62502492; in part by the Natural Science Foundation of Jiangsu Province Basic Research Program under Grant BK20250489; in part by the French government, through the 3IA Cote d'Azur Investments in the project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001; and in part by Toyota Motor Europe.

References

- [1] Mahmoud Ali, Di Yang, and François Brémond. Are visual-language models effective in action recognition? a comparative study. In *ECCVW*, 2024. 3
- [2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *ICCV*, 2021. 1, 6
- [3] Remi Cadene, Hedi Ben-younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *CVPR*, 2019. 3
- [4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE TPAMI*, 2019. 3
- [5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 1, 3, 8
- [6] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *ICCV*, 2021. 1
- [7] Liu Chunhui, Hu Yueyu, Li Yanghao, Song Sijie, and Liu Jiaying. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv:1703.07475*, 2017. 2, 6
- [8] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *CVPR*, 2019. 3
- [9] Rui Dai, Srijan Das, and François Bremond. Learning an augmented rgb representation with cross-modal knowledge distillation for action detection. In *ICCV*, 2021. 1, 3, 6, 7, 8
- [10] Rui Dai, Srijan Das, and Francois Bremond. Ctrn: Class-temporal relational network for action detection. In *BMVC*, 2021. 3
- [11] Rui Dai, Srijan Das, Luca Minciullo, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Pdan: Pyramid dilated attention network for action detection. In *WACV*, 2021. 1, 3, 6
- [12] Rui Dai, Srijan Das, Kumara Kahatapitiya, Michael Ryoo, and Francois Bremond. MS-TCT: Multi-Scale Temporal ConvTransformer for Action Detection. In *CVPR*, 2022. 1, 3, 5, 6, 8
- [13] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. *IEEE TPAMI*, 2022. 2, 3, 6
- [14] Rui Dai, Srijan Das, Michael S. Ryoo, and Francois Bremond. Aan: Attributes-aware network for temporal action detection. In *BMVC*, 2023. 1, 3, 6, 7
- [15] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *ICCV*, 2019. 3
- [16] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living. In *ECCV*, 2020. 3
- [17] Srijan Das, Rui Dai, Di Yang, and Francois Bremond. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE TPAMI*, 2021. 3
- [18] Haodong Duan, Yue Zhao, Kai Chen, Dian Shao, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *CVPR*, 2022. 1
- [19] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, 2020. 1
- [20] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 1
- [21] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *CVPR*, 2021. 1
- [22] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *IJCNN*, 2005. 6
- [23] K. Hara, H. Kataoka, and Y. Satoh. Learning spatio-temporal features with 3D residual networks for action recognition. In *ICCVW*, 2017. 1
- [24] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE TPAMI*, 2013. 1
- [25] Kumara Kahatapitiya and Michael S. Ryoo. Coarse-fine networks for temporal activity detection in videos. In *CVPR*, 2021. 3, 6
- [26] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *CVPR*, 2017. 3
- [27] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 3
- [28] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACMMM*, 2022. 1
- [29] Lina Mezghani, Piotr Bojanowski, Karteek Alahari, and Sainbayar Sukhbaatar. Think before you act: Unified policy for interleaving language reasoning with actions. *arXiv:2304.11063*, 2023. 3
- [30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr

- Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023. 1
- [31] AJ Piergiovanni and Michael S Ryoo. Learning latent super-events to detect multiple activities in videos. In *CVPR*, 2018. 6
- [32] AJ Piergiovanni and Michael S. Ryoo. Temporal gaussian mixture layer for videos. In *ICML*, 2019. 6
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 3
- [34] Hanoona Rasheed, Muhammad Uzair khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Finetuned clip models are efficient video learners. In *CVPR*, 2023. 1
- [35] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. *IEEE TPAMI*, 2019. 3
- [36] Michael S. Ryoo, Keerthana Gopalakrishnan, Kumara Kahatapitiya, Ted Xiao, Kanishka Rao, Austin Stone, Yao Lu, Julian Ibarz, and Anurag Arnab. Token turing machines. In *CVPR*, 2023. 6
- [37] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3D human activity analysis. *CVPR*, 2016. 3
- [38] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019. 1, 3
- [39] Mustafa Shukor, Corentin Dancette, and Matthieu Cord. epalm: Efficient perceptual augmentation of language models. In *ICCV*, 2023. 3
- [40] Mustafa Shukor, Corentin Dancette, Alexandre Rame, and Matthieu Cord. Unified model for image, video, audio and language tasks. In *ICCVW*, 2023. 3
- [41] Mustafa Shukor, Corentin Dancette, Alexandre Rame, and Matthieu Cord. UnIVAL: Unified model for image, video, audio and language tasks. *Transactions on Machine Learning Research*, 2023. 3
- [42] Gunnar A. Sigurdsson, Gül Varol, X. Wang, Ali Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 2, 3, 6
- [43] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In *ACM MM*, 2020. 1
- [44] Jing Tan, Xiaotong Zhao, Xintian Shi, Bin Kang, and Limin Wang. PointTAD: Multi-label temporal action detection with learnable query points. In *NeurIPS*, 2022. 1
- [45] Praveen Tirupattur, Kevin Duarte, Yogesh Rawat, and Mubarak Shah. Modeling multi-label action dependencies for temporal action localization. In *CVPR*, 2021. 3, 6
- [46] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 1
- [47] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Ge Yuying, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. In *CVPR*, 2023. 3
- [48] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *CVPR*, 2021. 1
- [49] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *CVPR*, 2023. 1
- [50] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*, 2024. 1, 3
- [51] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *ICCV*, 2017. 6
- [52] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding. *arXiv preprint arXiv:2105.09996*, 2021. 3
- [53] S. Yan, Yuanjun Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI*, 2018. 1
- [54] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, 2023. 3
- [55] Di Yang, Rui Dai, Yaohui Wang, Rupayan Mallick, Luca Minciullo, Gianpiero Francesca, and Francois Bremond. Selective spatio-temporal aggregation based pose refinement system: Towards understanding human activities in real-world videos. In *WACV*, 2021. 3, 4, 6
- [56] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Unik: A unified framework for real-world skeleton-based action recognition. In *BMVC*, 2021. 1, 4
- [57] Di Yang, Yaohui Wang, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Lac - latent action composition for skeleton-based action segmentation. In *ICCV*, 2023. 2, 3, 6, 8
- [58] Di Yang, Yaohui Wang, Quan Kong, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Self-supervised video representation learning via latent time navigation. In *AAAI*, 2023. 1, 5
- [59] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Via: View-invariant skeleton action representation learning via motion retargeting. *IJCV*, 2024. 1, 2, 3, 4
- [60] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *IJCV*, 2018. 2, 6
- [61] Chuhan Zhang, Ankush Gputa, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *CVPR*, 2021. 3

- [62] Xiaoguang Zhu, Ye Zhu, Haoyu Wang, Honglin Wen, Yan Yan, and Peilin Liu. Skeleton sequence and rgb frame based multi-modality feature fusion network for action recognition. *ACM TOMM*, 2022. [1](#), [6](#), [7](#), [8](#)
- [63] Yuhan Zhu, Guozhen Zhang, Jing Tan, Gangshan Wu, and Limin Wang. Dual detr for multi-label temporal action detection. In *CVPR*, 2024. [1](#), [3](#), [6](#)