# MoVie: Broaden Your Views with Human Motion for Action Detection - Supplementary Material

Di Yang[1] Mahmoud Ali[2] Xuanlong Yu[3] Xi Shen[3] Quan Kong[4]
Gianpiero Francesca[5] François Brémond[2]
[1]Suzhou Institute for Advanced Research, University of Science and Technology of China, China
[2]Inria Center at Université Côte d'Azur, France
[3]Intellindust AI Lab, China [4]Woven by Toyota, Japan [5]Toyota Motor Europe, Belgium

## Appendix

In this supplementary material, we provide further details to enrich the experimental analysis presented in the main paper. Section A elaborates on the implementation of our framework. Section B provides additional quantitative comparisons, extended analyses, and qualitative results to demonstrate the effectiveness of MoVie in action detection tasks. The code of this work and experiments will be made publicly available.

## A. Implementation Details

This section provides additional implementation details of the key modules used in MoVie, including the visual encoder, the Structured Motion Projection (SMP) module, the Motion-guided Feature Regularization (MGFR) module, and the temporal processing pipeline. More results, computation cost, and analysis are provided in the supplementary material.

### A.1. Visual Feature Extractor

We adopt two visual backbones depending on the experiment protocol: (i) pre-trained I3D [1] for comparison with prior state-of-the-art, and (ii) ViCLIP [7], a video foundation model based on a ViT-style spatio-temporal encoder with CLIP text supervision. We keep the visual encoders frozen unless otherwise specified, and extract per-frame visual embeddings $\mathbf{F_v} \in \mathbb{R}^{C_v \times T}$.

### A.2. Structured Motion Projection

The architecture summary of SMP and MGFR modules for learning Motion-Video representation is shown in Tab. 1. Specifically, given a skeleton sequence, we extract motion features using a spatio-temporal network following UNIK [9]. This produces per-frame, per-person features $\mathbf{F_m}$ (abbreviated as $\mathbf{F}$):

$$\mathbf{F} = \{\mathbf{f}_t^{(p)} \mid t = 1...T,\, p = 1...M\}, \quad \mathbf{f}_t^{(p)} \in \mathbb{R}^{C_m}.$$

SMP projects these features onto a learnable dictionary of motion primitives

$$\mathbf{D_m} \in \mathbb{R}^{K \times C_m},$$

where $K = 128$. The primitive activation map is computed as:

$$\boldsymbol{\alpha}_{t,p} = \mathbf{D_m}\, \mathbf{f}_t^{(p)}.$$

To reduce noise from pose estimation, we pass the raw activations through a lightweight MLP:

$$\tilde{\boldsymbol{\alpha}} = \sigma(\boldsymbol{\alpha}),$$

and then apply multi-person pooling to obtain a unified per-frame structured descriptor:

$$\hat{\boldsymbol{\alpha}} = \mathcal{G}_p(\tilde{\boldsymbol{\alpha}}) \in \mathbb{R}^{K \times T},$$

where $\mathcal{G}_p$ is max/mean pooling after a linear projection.

### A.3. Motion-guided Feature Regularization

MGFR aligns visual features to structured motion features by projecting $\mathbf{F_v}$ onto the orthogonal basis $\mathbf{Q} \in \mathbb{R}^{K \times C_v}$:

$$\mathbf{F_{mv}} = \mathbf{F_v} + \mathbf{Q}^\top \hat{\boldsymbol{\alpha}}.$$

Orthogonality is enforced to ensure non-redundant and physically meaningful motion directions. MGFR ensures that motion primitives modulate visual features without collapsing semantic variance.

### A.4. Temporal Modeling

We follow MS-TCT [2] for temporal processing. At each iteration, the temporal encoder takes the concatenated features: cat$[\mathbf{F_{mv}}, \mathbf{F_h}]$, as input, where $\mathbf{F_h}$ is the cached history feature from the previous window. A sliding window of length 300/60/60/100 is used for TSU, Charades, Multi-THUMOS, and PKU-MMD. The output is fed into a classification head for frame-wise predictions.

| Stage | Operation / Output Shape |
|---|---|
| Input | Skeleton Sequence: $C_{in} \times T \times M \times J$ |
| Motion Extraction | Spatial MHA + Temporal TCN layers<br>Output: $\mathbf{F} \in \mathbb{R}^{C_m \times T \times M}$ |
| SMP (Primitive Projection) | $\boldsymbol{\alpha} = \mathbf{QF}$, $K = 128$ |
| MLP Refinement | $\tilde{\boldsymbol{\alpha}} = \sigma(\boldsymbol{\alpha})$ |
| Multi-person Pooling | $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^{K \times T}$ |
| MGFR | $\mathbf{F_{mv}} = \mathbf{F_v} + \mathbf{Q}^\top \hat{\boldsymbol{\alpha}}$ |
| History-aware Temporal Model | $\mathbf{F_{mvh}} = \mathrm{TM}(\mathrm{cat}[\mathbf{F_{mv}}, \mathbf{F_h}])$ |
| Classifier | FC + Sigmoid |
| Output | Frame-level action labels |

Table 1. Main components used in MoVie.

| **Feature Extractor** | **TSU CS (%)** |
|---|---|
| MoVie w/ CLIP [6] | 48.6 |
| MoVie w/ X-CLIP [5] | 49.6 |
| MoVie w/ ViCLIP [7] | **50.1** |

Table 2. Performance (mAP) on TSU-CS in analyzing the visual feature extractor (*i.e.*, visual encoder) of MoVie.

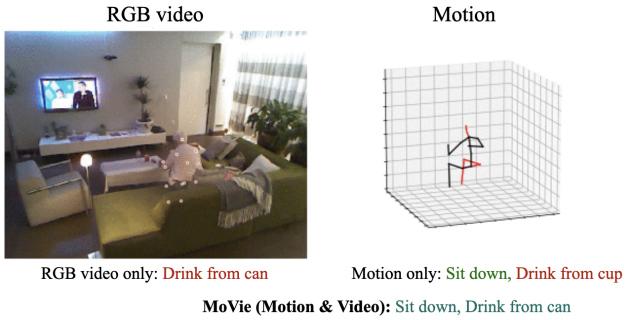


Figure 1. Qualitative results for action detection. In contrast to current action detection approaches using a single modality, **MoVie** takes a broad view that features two modalities (**video** and **motion**). MoVie achieves SoTA performance.

## B. Additional Results and Analysis

This section presents further quantitative and qualitative MoVie evaluations to analyze the settings targeting action detection tasks.

### B.1. Adaptation with Different Visual Extractors

A key advantage of MoVie is its ability to work as a plug-and-play motion adapter that can be attached to different visual backbones. To verify this property, we evaluate MoVie with several popular visual encoders, including CLIP [6], X-CLIP [5], and ViCLIP [7]. As shown in Tab. 2, MoVie consistently improves performance across all extractors.

Among them, ViCLIP provides the highest accuracy due to its large-scale video-language pretraining. X-CLIP and CLIP also yield competitive results, which confirms that MoVie does not rely on a specific visual backbone. The improvements indicate that structured motion offers complementary cues that remain effective even when the visual encoder varies in capacity or training data. This flexibility suggests that MoVie can adapt to future stronger vision models without requiring architectural changes.

### B.2. Analysis on Temporal Modeling

We first evaluate how different ways of adding history information influence temporal modeling. As shown in Tab. 3, simple feature concatenation provides the best accuracy. Attention-based fusion gives a small gain, but concatenation remains more stable in online inference. We also compare pure TCNs with the ConvTransformer structure. The combined model performs slightly better, which is consistent with earlier findings that TCNs capture short-term motion while Transformers model longer-range dependencies. These improvements are modest but confirm that our temporal module works reliably with different history and architecture choices.

### B.3. Temporal Window Size

We study the effect of different sliding-window sizes used during training and online inference. Results in Tab. 4 show that the best window size depends on dataset characteristics. Charades benefits from a shorter window due to dense activity changes, while TSU performs best with a larger window because of its long untrimmed sequences. This analysis helps guide practical parameter selection but does not affect the core design of our method.

### B.4. Comparison of 2D and 3D Skeletons

We compare the performance of MoVie using 2D and 3D skeleton inputs on TSU. Although 3D skeletons contain richer structural information, they may be less reliable in real-world scenes because current 3D pose estimators often fail under occlusion and extreme viewpoints [8]. In con-

| History Features | TSU CS (%) |
|---|---|
| w/o history | 48.1 |
| w/ attention | 49.3 |
| w/ concatenation | **50.1** |
| TCNs | 49.4 |
| ConvTransformer | **50.1** |

Table 3. mAP on TSU-CS and Charades in analyzing Temporal Modeling.

| Window Size | TSU CS (%) | Charades (%) |
|---|---|---|
| $W_s$=30 | 47.7 | 31.4 |
| $W_s$=60 | 49.4 | **33.5** |
| $W_s$=100 | 50.0 | 30.8 |
| $W_s$=300 | **50.1** | 24.3 |
| $W_s$=500 | 46.0 | 16.2 |

Table 4. Performance (mAP) on TSU CS and Charades in analyzing observed temporal sliding window size.

| Motion Data | TSU CS (%) | TSU CV (%) |
|---|---|---|
| 3D Skeletons | 46.3 | 29.5 |
| 2D Skeletons | **50.1** | **30.1** |

Table 5. Performance (mAP) on TSU CS and CV in analyzing the 2D/3D skeleton data.

trast, 2D skeleton estimation is generally more stable [4].

As shown in Tab. 5, 2D skeletons achieve better performance in the cross-subject setting and perform on par with 3D skeletons in the cross-view setting. This suggests that reliable motion cues matter more than dimensionality. The results confirm that MoVie adapts well to different types of skeleton data. We believe that high-quality 3D poses (e.g., from MoCap systems) could further boost performance but are not required for the effectiveness of our approach.

### B.5. Application with Noisy Skeletons

MoVie is designed to remain effective even when the skeleton modality is imperfect. The datasets used in our evaluation contain real-world challenges such as occlusion, viewpoint changes, and lighting variation, yet the estimated skeletons still provide useful motion cues. To quantify robustness, we inject Gaussian noise into joint coordinates and report the results in Tab. 7. The performance decreases by less than 2% even with relatively large perturbations, suggesting that the structured motion representation and MGFR alignment are stable under noisy inputs. This also shows the benefit of multi-person pooling, which reduces the impact of occasional pose estimation failures.

### B.6. Motion Representation and Design Choices

To show that the structured motion representation plays a critical role, we further provide the results using SMP alone

| Motion Representation and Design Choices | | |
|---|---|---|
| **Setting** | **TSU CS (%)** | **Charades (%)** |
| SMP-only | 32.6 | 19.1 |
| MoVie (full) | **50.1** | **33.5** |
| Random Init. | 46.2 | 30.0 |
| PCA | 46.4 | 31.1 |
| Pretrained (Ours) | **50.1** | **33.5** |
| $K$=16 | 39.7 | 28.1 |
| $K$=160 | 49.9 | 33.3 |
| $K$=128 | **50.1** | **33.5** |

Table 6. Additional ablation results on TSU CS and Charades.

| Noise Level (std) | TSU-CS (%) | Charades (%) |
|---|---|---|
| No noise | 50.1 | 33.5 |
| $\sigma = 2$ | 49.1 | 32.4 |
| $\sigma = 4$ | 48.6 | 31.7 |

Table 7. Analysis on MoVie when applied with noisy skeletons on TSU-CS and Charades.

(see Tab. 6). It confirms that motion primitives are most effective when combined with visual features. We also compare different dictionary initializations and observe that the pretrained motion dictionary consistently outperforms random or PCA initialization, indicating that meaningful motion primitives are essential for accurate motion–visual alignment. We further evaluate different dictionary sizes and find that performance remains stable when the number of primitives is sufficiently large, while too small dictionaries limit motion expressiveness.

### B.7. Computational Expense

We report the training cost and FLOPs of MoVie in Tab. 8. All models are trained on 4 $\times$ V100 GPUs with a window size of 300 frames on TSU. The results show that MoVie introduces moderate computational overhead compared with single-modal RGB or motion models, but remains significantly more efficient than previous multi-modal approaches.

MoVie achieves this balance because the Structured Motion Projection and MGFR modules operate on low-dimensional motion descriptors rather than dense feature maps. As a result, the added computation remains small relative to the visual backbone. The comparison confirms that MoVie is a practical plug-in module that improves detection accuracy while keeping computational cost manageable.

### B.8. Qualitative Results

To further illustrate the effectiveness of MoVie compared to single modality methods [2, 10], we include examples of their action detection performance on real-world TSU videos (see Fig. 1 and the attached video). These examples showcase the proposed MoVie, which can accurately

| Methods | Modality | Training Time (hrs) | FLOPs (G per window) |
|---|---|---|---|
| MS-TCT [2] | RGB | 19.5 | 6.6 |
| SD-TCN [3] | Motion | 17.0 | - |
| LAC [10] | Motion | 21.5 | - |
| MMFF [11] | Motion+RGB | 48.0 | 10.7 |
| **MoVie (Ours)** | Motion+RGB | 33.0 | 9.5 |

Table 8. Training cost and FLOPs comparison on TSU-CS. FLOPs are computed per sliding window of 300 frames.

detect complex actions in untrimmed videos, even in challenging scenarios involving multiple persons and compositional (overlapping) activities. MoVie is more robust to distractor objects (*e.g.*, book and TV) and at the same time motion-aware to human motion. The results demonstrate the potential of MoVie for real-world applications such as healthcare activity monitoring and human-computer interaction.

## References


[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 1

[2] 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, 4

[3] 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. 4

[4] Haodong Duan, Yue Zhao, Kai Chen, Dian Shao, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *CVPR*, 2022. 3

[5] 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. 2

[6] 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. 2

[7] 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, 2

[8] 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. 2

[9] 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

[10] 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. 3, 4

[11] 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. 4