InternVideo is Fine-Grained: Resolving Temporal Aliasing of Sub-Second Actions

Cleveland Crossling[0009-0007-1425-0325] and
Dustin van der Haar[0000-0002-5632-1220]

Academy of Computer Science and Software Engineering, University of Johannesburg, Cnr University Road and Kingsway Avenue, Auckland Park, Johannesburg, Gauteng, 2092
clevelandc@uj.ac.za, dvanderhaar@uj.ac.za

Abstract. While Transformer-based architectures have saturated performance on standard Temporal Action Localization (TAL) benchmarks, they exhibit severe degradation on fine-grained datasets. This work posits that this performance gap is not a failure of downstream localization architectures, but a fundamental "resolution collapse" in upstream feature extraction, where fixed temporal strides introduce destructive aliasing on sub-second actions. To isolate this mechanism, the AnnChor260 dataset is utilized as a resolution stress test, characterized by an average action duration of 0.66 seconds, which sits at the theoretical limit of standard sampling pipelines. Experiments reveal a definitive performance inversion, where simple per-frame spatial extractors (Frame-ResNet50) significantly outperform specialized spatiotemporal baselines (I3D) by over 25% mAP (31.24% vs. 5.87%), confirming that for sub-second activities, temporal aliasing outweighs the benefits of motion context. Furthermore, while granular foundation models (InternVideo) achieve saturation-level performance (90.82% mAP), ablations demonstrate that this success is strictly a function of token density rather than model capacity. Increasing the temporal stride of the same model from 1 to 32 frames per feature results in a catastrophic drop to 28.80% mAP. These findings establish a new boundary condition for video representation learning, identifying that high-frequency feature extraction, independent of parameter scale, is a requisite for robust sub-second localization.

1 Introduction

Temporal Action Localization (TAL) is a foundational component of video understanding. TAL datasets such as Thumos-14 [1] and ActivityNet [2] have seen substantial performance gains in metrics since the introduction of transformers into traditional localization methods. By modeling long-range dependencies through self-attention, TAL models have achieved performance saturation on these datasets, yielding precise temporal boundary delineation and robust action classification. Consequently, prevailing paradigms rely heavily on temporally distinct and unambiguous features generated by upstream representations.

Recognizing this limitation, the field is increasingly shifting toward more realistic and complex datasets [3,4]. An important subset of these datasets is referred to as fine-grained, typically characterized by dense temporal annotations and a prevalence of short-duration actions within continuous video streams, alongside low inter-class variability [5]. State-of-the-art models show performance degradation on these complex datasets despite theoretically possessing the temporal resolution to capture the fine actions presented [4,6]. The divergence between the theoretical temporal resolution of these architectures and their empirical underperformance suggests that the shortcoming lies in the quality of the video feature representation, rather than the model architecture, for fine-grained localization.

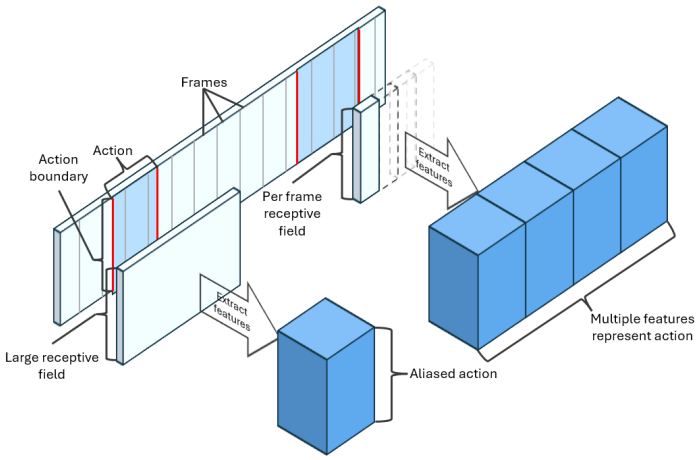

Fig. 1. The mechanism of temporal aliasing via receptive field aggregation. The top pathway illustrates a granular, per-frame receptive field where the sampling rate satisfies the representation requirements for sub-second actions, preserving distinct temporal boundaries. In contrast, the bottom pathway depicts a large temporal receptive field characteristic of standard spatiotemporal pooling. When the temporal receptive field exceeds the duration of the action, the feature extraction process undergoes a "resolution collapse," inevitably aggregating conflicting action and background semantics into a solitary, aliased embedding. This phenomenon acts as a low-pass filter, eliminating the high-frequency signal necessary for sub-second localization, regardless of the downstream architecture.

In this work, the existing methodologies for feature extraction are examined. It is hypothesized that aggressive temporal pooling significantly reduces the quality of fine-grained detail within the representation. The reduction in temporal resolution results in temporal aliasing of distinct action instances into solitary feature embeddings as seen in Figure 1, rendering boundary prediction ineffective. Consequently, extraction architectures designed to preserve high-frequency

features, such as SlowFast [7], are expected to yield improved metrics, regardless of the localization model employed.

To substantiate this hypothesis, three primary contributions are presented:

- **Empirical Demonstration of Temporal Inversion:** Identification of a phenomenon where pure spatial extractors outperform spatiotemporal baselines (31.24% vs 5.87% mAP), confirming that fixed strides induce destructive aliasing on sub-second actions.
- **Validation of Granular Temporality:** Validation of InternVideo as the requisite solution for fine-grained localization, achieving 89.83% mAP and overcoming the resolution collapse inherent to fixed-stride convolutional architectures
- **Definition of the Temporal Resolution Limit:** Establishment of a resolution stress test using AnnChor260 to isolate the specific failure mode of extraction pipelines at sub-second average action durations.

2 Related Work

2.1 Temporal Action Localization

The introduction of Transformer architectures [8] has established a new state-of-the-art for TAL. The ability to contextualize semantic features over extended temporal windows is especially important in localizing ambiguous action boundaries.

ActionFormer. ActionFormer [9] is a single-stage model that leverages a Transformer-based encoder to establish temporal context. It uses local self-attention for efficient dependency modeling and employs a feature pyramid to handle actions of varying scales.

TriDet. It is characterized by its key innovation, its Trident-Head [10], which is responsible for predicting the relative probability distribution around action boundaries. The Trident mechanism specializes in mitigating boundary ambiguity prevalent in most TAL applications.

AdaTAD. AdaTAD focuses on robustly representing action proposals over diverse durations [11]. It uses an Adaptive Template-Guided Self-Attention Network to generate proposal representations. This design is specialized for proposing temporal Intersection over Union (IoU) maps over a two-level refinement scheme.

ActionMamba. A more recent development, ActionMamba [12] is a linear complexity alternative to Transformer models for TAL. It replaces Transformer blocks with Mamba State Space Model (SSM) blocks, which enables efficient long-range dependency modeling with linear complexity.

Despite the significant architectural diversity of state-of-the-art detectors, the reliance on high-fidelity pre-extracted video representations remains a unifying constraint. Consequently, even the most advanced sequence model cannot recover the temporal information lost during extraction, imposing a theoretical upper bound on the accuracy of action-boundary localization.

2.2 Video Representation Learning

Feature extraction constitutes the foundational stage for downstream localization, imposing a fundamental constraint on system performance. To delineate between feature extraction mechanisms, each paradigm is analyzed from a balanced perspective that considers both its spatial and temporal representations.

Pure Spatial Extractors. Spatial extractors derive features by collating image representations of each video frame. Notably, while often powerful for individual image downstream tasks, no temporal information is propagated between features, resulting in a weaker spatiotemporal signal. Notable spatial models include DINOv3 [13] and ResNet50 [14].

Fixed Stride Spatiotemporal Modeling. The introduction of I3D [15] and S3D [16] represented a paradigm shift by expanding 2D kernels into 3D, allowing for the simultaneous capture of motion and appearance. While these models theoretically capture temporal context, standard implementations rely on fixed temporal strides. Aggressive pooling on the temporal axis is hypothesized to introduce aliasing, acting as a bottleneck for fine-grained boundary precision.

Multi-Pathway Architectures. To address the emerging presence of sub-actions in datasets, a characteristic of fine-grained datasets, SlowFast [7] was proposed as a dual-pathway solution. SlowFast consists of two distinct pathways: A *Slow* pathway, which extracts a dense representation of spatial information at a low frame rate, and a *Fast* pathway responsible for capturing nuanced motion information at a higher temporal resolution. While this multi-pathway architecture is useful for representing most fine-grained datasets, it is theoretically bounded by the temporal stride of the spatial pathway. This work suggests that motion lacks context, even if represented at high temporal fidelity, if the semantic spatial information is temporally sparse.

Granular Temporality Extractors. Originally, action recognition datasets were characterized by distinct trimmed video clips with singular, plainly represented actions. Early models excelled at classifying these actions as they were

semantically unambiguous over a short temporal duration. Modern foundation models such as InternVideo2-Giant [17] and its predecessors function as robust clip representation extractors, exploiting fine-grained temporality by encoding discrete video segments of variable durations. Each segmented clip is represented through a powerful spatiotemporal backbone. While highly effective, computation becomes substantially more demanding as granularity becomes finer.

3 Experiment Setup

3.1 Datasets

Current benchmarks in Temporal Action Localization have progressively shifted from sparse, distinct events to include denser, more fine-grained activities. As highlighted in recent literature, datasets such as MultiTHUMOS, FineAction, and FineGym [4, 18, 19] have established the standard for high-density localization. However, while these benchmarks introduce semantic complexity, they do not sufficiently stress the fundamental temporal resolution limits of feature extraction architectures.

Table 1. Comparison of temporal granularity and action density across standard TAL benchmarks. AnnChor260 exhibits a significantly lower Average Action Duration (AAD) compared to existing fine-grained datasets.

Dataset	Tot. Act	Avg. No. Act	Avg. Dur	AAD
FineAction [4]	103,324	6.17	-	7.1s
MultiTHUMOS [18]	38,690	93.6	-	3.6s
FineGym [19]	32697	-	-	1.7s
MultiSports [20]	37701	11.8	20.9s	1.0s
AnnChor260 [6]	25600	25.9	75.65s	0.66s

Analysis of these existing fine-grained datasets reveals a critical gap in temporal granularity. As detailed in Table 1, standard benchmarks, such as FineAction and MultiTHUMOS, exhibit average action durations of approximately 7.1 seconds and 3.6 seconds, respectively. While semantically challenging, these durations remain well within the receptive field of standard pooling operations; a standard I3D implementation with a stride of 16 at 25 fps covers approximately 0.64 seconds ($0.64s = 16/25$). Consequently, existing SOTA methods can resolve these boundaries without explicitly preserving high-frequency features, as the action duration significantly exceeds the sampling period.

To evaluate the resolution collapse hypothesis, this study utilizes the AnnChor260 dataset [6] as a definitive stress test. Unlike its predecessors, AnnChor is characterized by an average action duration of only 0.66 seconds and a high action density of 25.9 instances per video. Notably, this sub-second average duration aligns with the theoretical limit of standard extraction pipelines. By concentrating the experimental ablation on AnnChor, we isolate the specific failure

mode of temporal aliasing, utilizing this dataset as a predictive proxy for model performance in the limit of temporal granularity.

3.2 Feature Extraction Methodologies

To validate the impact of temporal resolution on localization performance, a set of extractors is selected to contrast fixed-stride architectures against variable-granularity sampling models.

Standard Baselines As a control for standard spatiotemporal modeling, I3D [15] and SlowFast [7] are employed. These models represent the prevailing standard for generating TAL features. As I3D is the most common prevailing standard for video feature extraction for most applications, it is specifically utilized as the baseline for results gathering. To validate the significance of action aliasing and to provide a baseline for pure-spatial representations, Frame-ResNet50 [14], and Frame-DINOv3 [13] are evaluated.

Feature Fusion To address the hypothesized aliasing limitations of standard baselines, a concatenation representation of DINOv3 [13] and SlowFast [7] is created. The concatenation of features is performed by nearest neighbor matching, with temporal features to their respective spatial frames. The result of the temporal matching mechanism is a non-interpolated "staircase" relationship between spatiotemporal features. This feature fusion theoretically permits the temporal context of SlowFast to augment the high-fidelity spatial representations extracted by DINOv3.

InternVideo2 Fixed-stride pooling is theorized as the primary adversary to sub-action localization, especially when action duration is sub-second. To address this limitation, InternVideo2 [17] is employed as the primary representative for granular temporality extractors. Unlike I3D, InternVideo2 operates on discrete video clips, allowing for the dense sampling of variable-duration segments without architectural aliasing.

Two instances of InternVideo2 are utilized to determine the effect of parameter increase on TAL performance. These two versions of InternVideo2 are as follows: The one billion parameter (1B) version and the six billion parameter (6B) version. Both versions utilize the same underlying masking architecture (VideoMAE) to create discrete video clips, which are then extracted into video features. The major differentiator is the quantity of parameters for each model, independent of training data and structure [17].

3.3 Parameters

To ensure that the results gathered remain constrained to a control, a set of parameters is specifically selected. All extracted features are gathered at the

minimal temporal stride or on a per-frame basis, depending only on available model versions (e.g., SlowFast 8×8 and SlowFast 16×8 [7]). ActionFormer [9] is utilized as the model for the downstream task and is prepared with the following configuration:

1. Number of Classes: 11, consistent with AnnChor260.
2. Input Dimension: Variable depending on input feature dimension.
3. Frame Rate: 25 fps, consistent with the AnnChor dataset specification.
4. Feature Stride: Dynamically adapted to match the input stride of the extracted features.
5. Learning Rate: 1×10^{-4} with a warm-up and cosine annealing schedule, following the standard ActionFormer protocol [21].
6. Epochs: 35.
7. Max Sequence Length: 4480 tokens, sufficient to cover the longest video in the dataset without truncation.
8. Multi-Head Attention Window Size: Fixed at 20 tokens, specifically selected to optimize for the dataset's short average action duration.

All other parameters are consistent with the THUMOS-14 default configuration provided by the ActionFormer repository [22].

3.4 Measure of Success

Success is evaluated primarily using mean Average Precision (mAP) averaged across temporal Intersection over Union (tIoU) thresholds (0.3, 0.5, and 0.7). However, the primary indicator of success is the performance gain between standardized extraction techniques and dense, per-frame methods.

On the AnnChor dataset, if dense methods provide a significant improvement, it may suggest that the hypothesis of this work is correct. Actions are being aliased by temporal strides that are too long for sub-second actions to be adequately represented. A negligible increase would suggest that temporal resolution does not govern the representation quality of fine-grained actions.

4 Results

This section displays the results gathered from the experiment setup.

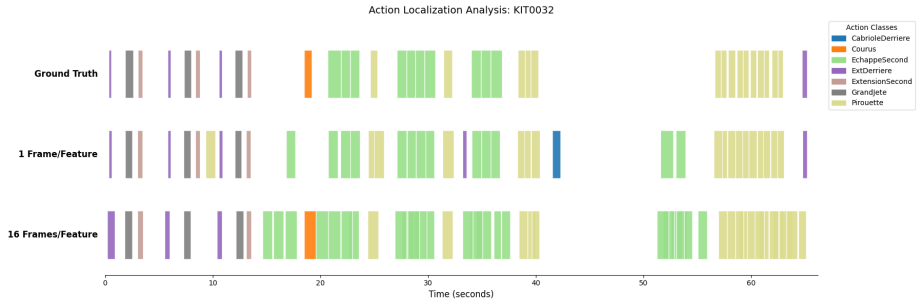

Fig. 2. Qualitative comparison of temporal action localization on a representative AnnChor-260 sequence. The top row displays the ground truth, characterized by dense, short-duration actions. The middle row (InternVideo-1B, 1 frame/feat) demonstrates robust boundary recovery, effectively resolving individual instances. In contrast, the bottom row (InternVideo-1B, 32 frames/feat) exhibits severe **temporal aliasing**, where distinct actions are fractured into prediction artifacts due to the loss of high-frequency signal during feature extraction.

tIoU	I3D	I3D-S	SF	F-Res	F-DINO	DINO-SF	Int.Vid.1B	Int.Vid.6B
0.3	9.24	12.00	37.05	27.64	38.65	41.88	91.41	92.54
0.5	5.78	7.90	30.35	22.09	31.63	33.51	90.43	91.67
0.7	2.26	3.37	17.85	14.11	23.10	22.13	81.24	82.51
Ave. mAP	5.87	7.64	29.13	21.29	31.24	32.83	88.43	89.83

Table 2. Comparative Evaluation of Feature Extraction Architectures on the AnnChor260 Benchmark. Performance is reported as mean Average Precision (mAP) across temporal Intersection over Union (tIoU) thresholds of {0.3, 0.5, 0.7}.

Ablation Method	Average mAP (%)
1B 1f/feat (Default)	88.45
6B 1 f/feat	89.83
1B 2 f/feat	88.39
1B 4 f/feat	88.08
1B 8 f/feat	85.35
1B 16 f/feat	67.92
1B 32 f/feat	28.80
1B 45 Epochs	90.82

Table 3. Ablation Study of InternVideo Architecture on AnnChor260. The impact of temporal stride, parameter scale, and training duration on localization performance is analyzed.

5 Discussion

5.1 Spatial Granularity over Temporal Context

A critical and counterintuitive finding of this study is the significant performance inversion between pure spatial extractors and standard spatiotemporal baselines. Specifically, as detailed in Table 4 Frame-ResNet50 (21.29% mAP) and Frame-DINOv3 (31.24% mAP) significantly outperform the I3D baseline (5.87% and 7.64% segmented mAP), surpassing even the leading benchmark in the original AnnChor thesis (11.09% mAP on AnnChor260) [6]. Standard video understanding paradigms posit that motion dynamics are essential for action discrimination. However, these results indicate that for sub-second actions, temporal resolution is the dominant bound on performance.

This inversion is attributed to the structural limitations of hierarchical temporal pooling. Standard spatiotemporal architectures, such as I3D, utilize progressive downsampling layers that compress the temporal dimension, typically compressing 16 frames into a singular feature vector [15]. This aggregation mechanism introduces destructive aliasing when applied to sub-second activities, as shown in Figure 2. In contrast, pure spatial extractors and the InternVideo foundation model circumvent this bottleneck by maintaining a high token density [17, 23], thereby preserving the signal frequency necessary to resolve short-duration events.

5.2 Combined-Context Feature Representation

To bridge the gap between standard spatiotemporal techniques and pure-spatial models, the extracted features of both pure-spatial and spatiotemporal models are aggregated. Specifically, Frame-DINOv3 and SlowFast features were combined via a naive nearest-neighbor upsampling mechanism to ensure dimension

matching. As detailed in Table 4, this combined-context representation yielded marginal gains (32.83% mAP), representing only a 1.59% mAP improvement over the pure Frame-DINOv3 baseline. While the aggregation method is comparatively simple, it demonstrates that the performance advantage of per-frame sampling cannot be replicated by supplementing it with pooled temporal resolution features.

5.3 Computational and Semantic Limitations of InternVideo

While InternVideo2 successfully resolves the temporal aliasing bottleneck, this capability incurs a substantial computational penalty that limits its deployability. The 6B variant requires magnitudes more computation time for inference than standard baselines such as I3D or ResNet50, rendering it intractable for real-time applications without significant distillation. Furthermore, despite its high mAP, InternVideo is limited by high-velocity changes such as jumps or blurs due to its VideoMAE backbone [17]. Considering the AnnChor dataset contains high-dynamic ballet movements, the impact of this limitation may be mitigated by annotation bias, where blurry transition frames are systematically excluded from ground-truth segments, effectively insulating the model from its specific weakness in reconstructing high-frequency temporal noise.

5.4 Dataset Generalization

This study identifies a critical "resolution collapse" on the AnnChor260 benchmark. It is acknowledged that the scope of this inversion phenomenon is constrained by the specific temporal density of the dataset. AnnChor260 was selected specifically because its average action duration sits below the theoretical effective sampling rate of standard fixed-stride architectures. If this methodology is applied to established fine-grained benchmarks such as FineGym [19], which exhibits longer average durations, they may not exhibit the same catastrophic degradation for spatiotemporal baselines.

6 Conclusion

This study establishes resolution collapse as the fundamental limit of current Temporal Action Localization architectures on fine-grained benchmarks. By utilizing the AnnChor260 dataset as a high-frequency stress test, this work demonstrates that the prevailing paradigm of fixed-stride spatiotemporal pooling induces destructive aliasing on sub-second actions. While the InternVideo foundation model effectively resolves this bottleneck to achieve remarkable performance, our ablation confirms that this success is strictly a function of its dense token preservation rather than model capacity alone. Consequently, these findings define a necessary trajectory for video representation: to adequately localize the sub-second frontier, future architectures must decouple temporal resolution from computational depth, prioritizing efficient signal granularity over indiscriminate parameter scaling.

References

1. Haroon Idrees, Amir R. Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, February 2017.
2. Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
3. Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization, 2019.
4. Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE Transactions on Image Processing*, 31:6937–6950, 2022.
5. Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1194–1201, 2012.
6. Margaux Bowditch and Dustin van der Haar. Annchor: A video dataset for temporal action localization in classical ballet choreography. In *Pattern Recognition*, pages 194–209, Cham, 2025. Springer Nature Switzerland.
7. Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition, 2019.
8. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.
9. Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, volume 13664 of *LNCS*, pages 492–510, 2022.
10. Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling, 2023.
11. Shuming Liu, Chen-Lin Zhang, Chen Zhao, and Bernard Ghanem. End-to-end temporal action detection with 1b parameters across 1000 frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18591–18601, June 2024.
12. Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Zhe Chen, Zhiqi Li, Jiahao Wang, Kunchang Li, Tong Lu, and Limin Wang. Video mamba suite: State space model as a versatile alternative for video understanding, 2024.
13. Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025.
14. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
15. Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017.

16. Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018.
17. Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning, 2022.
18. 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. *International Journal of Computer Vision*, 2017.
19. Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
20. Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13536–13545, October 2021.
21. Actionformer repository. https://github.com/BharathChalla/ActionFormer. Accessed: 2025-11-03.
22. Haroon Idrees, Amir Zamir, Yu-Gang Jiang, Alexander Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155, 04 2016.
23. Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.