Self-Supervised Hierarchical Motion Learning for Egocentric Activity Recognition in Long Videos

Claudio Smith
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
smith1980@uab.edu

Abstract

Egocentric activity recognition from long, uncurated video streams constitutes a foundational challenge for next-generation assistive technologies, augmented reality interfaces, and large-scale behavioral analytics. This paper presents a systems-level investigation into self-supervised hierarchical motion learning as a principled framework for addressing the temporal scale, viewpoint variability, and computational constraints inherent to egocentric video understanding. We conceptualize motion as a multiscale structural primitive and propose a hierarchical architecture that decomposes egocentric action into layered motion patterns, from fine-grained hand-object interactions to coarse daily routines. The self-supervised learning paradigm leverages the natural temporal coherence and cross-scale correlations of first-person footage, eliminating the prohibitive cost of dense manual annotation. Our discussion centers on system architecture, infrastructure trade-offs, and the integration of such models into real-world socio-technical pipelines. We analyze memory-compute co-design strategies for on-device and cloud-edge deployment, examine the robustness of learned representations under domain shift and egocentric sensor noise, and articulate the fairness risks embedded in training data distributions collected from narrow demographic profiles. Furthermore, we map out governance and policy implications, including the tension between behavioral inferencing accuracy and privacy preservation, and propose transparency mechanisms for model auditing. Through cross-domain comparisons with third-person video understanding systems and large-scale language-vision models, we highlight the unique structural demands of egocentric motion hierarchies. The paper concludes with a forward-looking agenda that integrates sustainability metrics, regulatory alignment, and participatory infrastructure design, arguing that hierarchical self-supervised motion learning is not merely a technical architecture but a bridge between low-level perception and ethically grounded, long-term activity intelligence.

Keywords

egocentric activity recognition, self-supervised learning, hierarchical motion representation, long video understanding, system architecture, infrastructure, fairness, governance, sustainability.

1. Introduction

The proliferation of wearable cameras and head-mounted devices has produced an unprecedented volume of egocentric, or first-person, video data. These long, multi-hour recordings capture daily life from the wearer's perspective, presenting a radically different structure from the short, curated clips that dominate third-person vision benchmarks. Recognizing activities from such streams is essential for applications ranging from memory augmentation for aging populations to automated skill assessment in industrial settings and

continuous health monitoring. However, the unique properties of egocentric vision—unbounded duration, extreme viewpoint motion, object interaction at close range, and the absence of the actor's own body in full view—defy traditional supervised recognition pipelines that depend on strongly labeled temporal boundaries and large annotated datasets. Manual annotation of long egocentric videos is prohibitively expensive, inconsistent, and does not scale, creating a pressing need for self-supervised approaches that can learn from the structural regularities inherent in the data itself. Simultaneously, deploying these models at scale demands a careful articulation of the underlying system architecture, the computational infrastructure that sustains learning and inference, and the broader socio-technical implications that emerge once activity recognition moves from the laboratory into everyday life.

Self-supervised learning has recently transformed natural language processing and static image understanding by deriving supervisory signals from the data itself, such as predicting masked portions of an input or contrasting augmented views of the same instance. Extending these paradigms to video, particularly to long egocentric sequences, requires grappling with the temporal dimension as a first-class citizen. Motion becomes a particularly rich source of self-supervision in egocentric video, because the wearer's own head movements, hand trajectories, and object manipulations encode intentional structure across multiple timescales. Rather than treating motion as a single optical flow field or a short-range feature, a hierarchical perspective recognizes that activities can be decomposed into a cascade of motion primitives: rapid, fine-grained finger articulations give way to wrist and arm trajectories, which in turn compose mid-level action segments, eventually aggregating into high-level daily routines such as preparing a meal or commuting to work. The challenge is to design a learning architecture and a corresponding system that can capture these nested temporal dependencies without explicit segmentation boundaries, while remaining computationally tractable for multi-hour input streams and robust to the distributional variability encountered across users, cultures, and environments.

This paper offers a broad, interdisciplinary analysis of self-supervised hierarchical motion learning for egocentric activity recognition in long videos, emphasizing not only the algorithmic principles but also the systems, infrastructure, governance, and policy dimensions that condition its successful adoption. We structure our discussion around the hierarchical decomposition of motion as an organizing framework, interrogating its architectural expression, the self-supervised objectives that sustain its training, and the performance-stability trade-offs that arise when embedding such models into real-world socio-technical ecosystems. The work draws from computer vision, distributed systems, human-computer interaction, and technology policy to provide a unified account of why hierarchical motion understanding represents a critical inflection point for egocentric AI and what structural choices must be made to align it with values of fairness, robustness, sustainability, and accountability.

2. Self-Supervised Hierarchical Motion: Conceptual Foundations

Understanding motion in egocentric video necessitates moving beyond monolithic representations that collapse temporal structure into a single feature vector. The hierarchical view posits that activities are not flat sequences but rather organized along spatial and temporal scales that mirror the body's own kinematic chain and the cognitive segmentation of goal-directed behavior. At the lowest level, rapid optical flow patterns and differential image signals correspond to the articulation of fingers, subtle head stabilization movements, and the

deformation of manipulated objects. These signals aggregate at an intermediate level into motion phrases—repeated gestures, transport motions, and object interaction maneuvers—that span a few seconds. At the highest level, those phrases compose extended activities such as cooking a dish or assembling furniture, which themselves form episodes of daily routines lasting minutes to hours. A hierarchical motion learning system must therefore construct representations that are simultaneously sensitive to sub-second detail and capable of integrating information across long temporal windows without succumbing to the quadratic complexity of dense attention mechanisms over tens of thousands of frames.

This multiscale requirement invites a system design that interleaves motion encoding at multiple resolutions, akin to a pyramid of feature streams. Rather than extracting motion as a post-hoc descriptor, hierarchical models embed motion computation directly into the representation backbone, enabling the network to learn what constitutes informative movement at each scale. Self-supervision emerges naturally in this setting because the causal structure of egocentric video provides pervasive weak labels: the future motion field is partially predictable from the current and recent past observations at a coarser scale, while fine-scale motion deviations often signal object contact events that are highly discriminative. By training a model to predict the coarse motion trajectory of the head and hands, or to contrast segments that share the same underlying activity against distractor segments from different portions of the video, the system can acquire a structured motion embedding without any human annotation. Such an approach reframes the traditional boundary between motion estimation and activity recognition, making motion not just an intermediate cue but the primary organizational principle for long-range temporal understanding.

Central to this conceptualization is the notion of motion as a latent hierarchical grammar. Drawing analogies from linguistic syntax, low-level motion tokens can be seen as phonemes that combine into morpheme-like short actions, which in turn form sentence-like activity clauses, ultimately producing a discourse-level narrative of the day. Although we explicitly avoid formal grammatical models or equation-based specifications, the architectural implication is clear: the system must support a hierarchy of temporal receptive fields that can be composed via learned, rather than hand-crafted, aggregation steps. This raises immediate infrastructure challenges, because training such a system on hundreds of hours of egocentric footage demands memory-efficient streaming architectures that can keep long-term motion context alive without storing all frame-level activations. The following sections map these conceptual pressures onto concrete architectural and infrastructural choices, examine the self-supervised learning objectives that animate the hierarchy, and evaluate the robustness, fairness, and governance concerns that emerge upon deployment.

3. Architectural Design and Infrastructure Trade-Offs

Realizing a hierarchical motion learning system for egocentric long videos forces designers to make critical decisions about the backbone architecture and the computational substrate on which it runs. The tension between global temporal context and local precision is acute: long videos contain thousands of frames, yet the most informative motion signals are often spatially localized, such as a hand grasping a cup. Transformer-based architectures have shown promise for video understanding by applying self-attention across space and time, but standard dense attention over all tokens becomes infeasible for sequences of the length encountered in egocentric settings. Prior work on video transformers has introduced factorized space-time attention, but even these strategies require significant memory and can degrade when temporal spans extend beyond a few minutes. Hierarchical motion architectures

respond to this challenge by structuring the attention or pooling operators according to the inherent multiscale nature of motion, enabling the system to compress fine-grained motion information into progressively coarser summaries without losing semantic continuity.

One promising architectural direction organizes the network into interleaved streams that process motion at different temporal granularities. The system might maintain a fast-stream with high frame-rate input that captures hand-object dynamics and a slow-stream operating on a heavily temporally subsampled sequence to track room-level navigation and activity context. This design echoes the SlowFast paradigm originally developed for third-person action recognition, but adapted here specifically for the egocentric viewpoint, where the fast stream is often dominated by ego-motion and object interactions while the slow stream captures scene transitions and global routine structure. The hierarchical extension introduces additional intermediate streams, effectively forming a multiscale motion pyramid. Information flows upward from fine to coarse scales through lightweight cross-scale attention modules, and downward from coarse to fine scales through conditional modulation, allowing high-level activity goals to inform the interpretation of ambiguous local motions. Training such an architecture in a self-supervised manner requires designing auxiliary tasks that operate across scales, such as predicting the coarse future trajectory from fine-scale motion history or aligning event boundaries between streams.

The infrastructure demands of training and deploying hierarchical motion models on long egocentric videos are substantial, and the choice between cloud-centric, edge-federated, and fully on-device paradigms carries significant implications for latency, privacy, and energy consumption. Cloud-based training on large egocentric corpora can leverage thousands of accelerators but introduces data transfer bottlenecks and raises acute privacy concerns, as raw first-person video is among the most sensitive data modalities. Edge-cloud collaboration models attempt to mitigate this by performing initial feature extraction on a local device and transmitting only abstracted motion embeddings to a central server for further aggregation and contrastive learning. This split computation strategy preserves a degree of privacy but requires careful co-design of the feature hierarchy so that intermediate representations are both compact and semantically rich enough to support self-supervised objectives. Fully on-device deployment, appealing for assistive applications that must function without connectivity, demands aggressive model compression, quantization, and the use of temporal sparse inference schedules that only activate the fine-scale streams when significant object interaction is detected. The hierarchical nature of the motion representation naturally lends itself to such conditional computation, as coarse-scale streams can serve as gating signals that wake up finer-scale processing only when needed, thereby reducing the average power draw and extending battery life in wearable devices.

A further infrastructural consideration involves long-term memory management. To recognize an activity that spans several minutes, the system must retain relevant motion context beyond the limited window of any single transformer layer. Hierarchical systems often implement a form of memory consolidation in which coarse-scale features are cached in a recurrent or state-space mechanism that summarizes past events and gradually fades out stale information. The design of this persistent memory interacts closely with the self-supervised learning objective: the system can be trained to predict whether a current motion segment is consistent with the summarized past, creating an implicit anomaly detection capability that improves activity boundary discovery. Memory compression must be tuned to the cadence of human activities, avoiding both premature forgetting of temporarily interrupted tasks and excessive

retention that blurs the distinction between distinct episodes. Such tuning is highly dependent on the cultural and contextual rhythms of the target population, introducing fairness dimensions that will be explored later.

4. Self-Supervised Learning Paradigms and Training Strategies

The hierarchical motion representation is coupled with self-supervised objectives that circumvent the need for manually labeled temporal boundaries. The most influential paradigms that can be adapted to this setting include contrastive learning, masked motion modeling, and future trajectory prediction, each emphasizing different aspects of the motion hierarchy and imposing distinct training infrastructure requirements. Contrastive learning in the egocentric video domain constructs positive pairs from temporal windows that share the same underlying activity, while negative pairs are drawn from temporally distant segments or from other videos. By applying contrastive loss at multiple levels of the feature hierarchy, the model learns to align representations of motion fragments that are semantically similar but may differ in low-level dynamics due to viewpoint jitter or lighting changes. Crucially, the definition of a positive pair must respect the hierarchical structure: a brief hand movement and a longer cooking sequence should not be contrasted at the same granularity; rather, the system can use nested contrastive objectives that apply loose correspondence at coarse scales and tighter correspondence at fine scales.

Masked motion modeling, inspired by the success of masked autoencoders in images and short videos, offers an alternative that is particularly well-suited to hierarchical architectures. The model receives a heavily masked version of the input motion stream at several scales and is tasked with reconstructing the missing motion tokens from the visible context. In egocentric video, masking can be applied both spatially within frames and temporally across frames, forcing the model to reason about long-range dependencies to fill gaps that extend over seconds or even minutes. A hierarchical decoder reconstructs the missing motion progressively, first generating a coarse trajectory scaffold and then refining it with higher-frequency detail. This strategy is computationally efficient during pre-training because only the visible tokens are processed by the encoder, reducing the effective sequence length, yet it demands a decoder architecture that can synthesize plausible fine-scale motions consistent with physical constraints and object affordances. The self-supervised signal is implicit in the reconstruction error, which can be differentiated with respect to all parameters, enabling end-to-end training at scale.

Future motion prediction constitutes a third paradigm that links hierarchical motion learning with embodied AI and planning. Egocentric video is inherently predictive: the wearer’s head trajectory, hand movement, and object state evolve according to intentions that are partially observable from the visual history. A hierarchical prediction objective requires the model to forecast coarse-scale motion embeddings over the next several seconds, as well as the transitions between activity phases, while leaving the fine-scale restitution as a conditional generation challenge. This predictive framework aligns naturally with reinforcement learning and imitation learning pipelines, suggesting that self-supervised models pre-trained for motion prediction could serve as initialization for policy learning in assistive robotics. However, predictive training is more demanding than contrastive or masked approaches, as it requires the model to capture not only what is typical but also the multimodal possibilities of human action, a requirement that interacts with dataset diversity and fairness in subtle ways.

A combined training curriculum can leverage all three paradigms by staging pre-training phases that progressively increase the temporal horizon and abstraction level. Early phases

might employ short-range contrastive learning and masked token reconstruction on clips of a few seconds, while later phases introduce extended future prediction and cross-stream consistency losses that bridge fine and coarse scales. This curriculum mirrors the hierarchical structure itself and echoes the way humans learn to parse continuous activity: first attending to brief action fragments, then gradually building a repertoire of routines. Training at this scale requires a robust orchestration infrastructure that can manage data sharding, gradient accumulation over long temporal windows, and mixed-precision computation across heterogeneous hardware. System logs and checkpoints must be designed to track fairness metrics during pre-training, ensuring that the learned motion hierarchy does not drift toward demographic-specific patterns that would fail for underrepresented user groups.

5. Robustness and Domain Generalization

Egocentric activity recognition systems deployed in the wild face a formidable array of distribution shifts that test the robustness of hierarchical motion representations. Wearable camera footage exhibits extreme variability in illumination, with sharp transitions between indoor and outdoor environments, rapid camera rotations that blur frames, and occlusions caused by the wearer's own hands and clothing. Moreover, the motion signatures of the same high-level activity can differ substantially across cultures, occupations, and personal habits. A hierarchical architecture, by separating scale-specific processing, offers inherent inductive biases that can improve robustness: fine-scale object manipulation features may be less sensitive to global scene changes, while coarse-scale navigation patterns may be more stable across different kitchen layouts as long as the underlying activity structure is preserved. Nevertheless, without explicit robustness interventions, self-supervised models can latch onto spurious correlations, such as associating a particular room's color distribution with an activity rather than the motion trajectory itself.

Designing for robustness requires attention to data curation, augmentation strategies, and test-time adaptation mechanisms. The hierarchical motion representation can be hardened by applying scale-specific augmentations during self-supervised pre-training: for example, fine-scale streams may be subjected to strong color jittering and local spatial perturbations, whereas coarse-scale streams receive augmentations that simulate different walking speeds or viewpoint rotations. Such targeted augmentation teaches the model to decouple motion semantics from low-level appearance. Test-time adaptation can leverage the hierarchical structure by detecting when the temporal coherence across scales indicates a distribution shift—for instance, when the coarse stream predicts a room transition but the fine stream observes an unexpected hand motion—and using that discordance signal to trigger a lightweight fine-tuning phase on recent unlabeled samples. This continuous adaptation loop, however, must be governed by privacy-preserving constraints, as the fine-tuning data remains on the device and must not be uploaded in raw form.

The robustness conversation extends to sensor failures and hardware heterogeneity. Wearable cameras vary in resolution, frame rate, and field of view; head-mounted displays introduce eye-tracking data that can enrich motion cues but may be unavailable in simpler devices. A hierarchical motion learning system should expose well-defined interfaces between streams so that missing modalities can be gracefully handled by substituting a coarser motion estimate or relying on a learned prior. In multi-sensor fusion scenarios, the hierarchical pyramid can act as a temporal integrator that reconciles asynchronous data from an inertial measurement unit and a camera, with the coarse scale providing a stable frame of reference that aligns noisy IMU trajectories with visual ego-motion. Robustness, therefore, becomes a cross-layer

property that spans the hardware, the motion representation, the training procedure, and the runtime adaptation logic.

6. Fairness, Governance, and Policy Implications

The deployment of self-supervised hierarchical motion recognition in egocentric applications raises profound questions of fairness, accountability, and governance. Because these systems learn directly from unlabeled data, any demographic, cultural, or socioeconomic biases present in the training corpus are absorbed into the motion hierarchy without the corrective buffer that a carefully curated annotation process might provide. If the pre-training dataset predominantly features able-bodied individuals from high-income urban environments performing a narrow set of activities, the resulting model will systematically misinterpret the motion patterns of elderly users, people with motor disabilities, or individuals in rural settings where daily routines and object interactions differ markedly. The very strength of self-supervision—its ability to capture subtle statistical regularities—becomes a liability when those regularities reflect structural inequities rather than universal human activity structure.

Addressing fairness in this context calls for a multi-pronged strategy that includes participatory data collection, fairness-aware pre-training objectives, and model auditing frameworks. Rather than relying on passively collected video from a single source, dataset construction must actively sample from diverse geographic regions, occupational settings, and ability profiles, with informed consent mechanisms that go beyond blanket data usage agreements. At the algorithmic level, hierarchical motion models can be constrained with adversarial debiasing techniques applied at the coarse scale, where demographic factors such as walking gait or clothing style might otherwise leak into activity representations. More fundamentally, the hierarchical decomposition offers an opportunity to separate motion patterns that are biomechanically constrained from those that are culturally conditioned, enabling downstream applications to fine-tune only the culturally sensitive layers while freezing the core motion primitives. Such decomposition requires careful governance oversight to define which layers fall into which category, a decision that cannot be purely technical but must involve ethicists, domain experts, and affected communities.

Governance frameworks for egocentric activity recognition are still nascent. The vast amount of behavioral data extracted by motion models—in effect, a continuous log of everything a person does—creates surveillance risks that extend far beyond the immediate application. Policy interventions must address data retention limits, usage constraints, and the right to contest model inferences. Real-time activity recognition integrated into workplace monitoring systems, for instance, could be used to discipline employees based on inferred productivity patterns derived from coarse-scale motion trajectories, with no human-readable justification. To mitigate such risks, regulatory structures should mandate transparency reports that describe the motion hierarchy’s decision process in terms understandable to laypersons, perhaps through automatically generated textual narratives that map each scale of the hierarchy to a semantic activity description. The European Union’s Artificial Intelligence Act and similar regulatory initiatives are beginning to classify emotion recognition and biometric categorization as high-risk, and continuous activity recognition from wearable cameras should arguably receive comparable scrutiny, especially when used in employment, education, or insurance contexts.

The tension between the utility of motion hierarchies for health monitoring and the potential for intrusion illustrates the need for contextual integrity. An elderly care system that recognizes meal preparation and medication routines can support independent living, but the

same motion hierarchy should not be repurposed for targeted advertising without explicit re-consent. Technical architectures that embed hierarchical access control—allowing a caregiver to access only coarse-scale activity summaries and not fine-grained hand-object interaction logs—can partially address this tension, but they demand that the system's representation space is structured to facilitate such selective disclosure. This requirement, in turn, affects the design of the hierarchy itself: scales must be semantically separable enough that revealing one does not inadvertently leak information that could be used to reconstruct the other. The interplay between architectural design and policy compliance thus becomes a central design axis, not an afterthought.

7. Sustainability, Scalability, and Lifecycle Management

Long video understanding at scale has an undeniable environmental footprint. Training hierarchical motion models on thousands of hours of egocentric footage using large transformer backbones consumes substantial electricity, and the lifecycle impact of deploying these models on millions of wearable devices—each performing continuous inference—amplifies those costs. The hierarchical structure offers several lever points for improving sustainability. The gated conditional computation strategy described earlier directly reduces per-device energy consumption by idling fine-scale streams during periods of low activity. Furthermore, the multiscale nature of the hierarchy allows for dynamic resolution and frame-rate scaling based on available energy budgets: on a device nearing battery depletion, the system can fall back to coarse-scale processing only, providing a degraded but still useful activity summary. This graceful degradation aligns with the principles of green AI and enables a spectrum of operation points rather than a single all-or-nothing deployment.

At the training stage, sustainability can be advanced by designing pre-training curricula that maximize sample efficiency and leverage model recycling across tasks. The hierarchical motion representation learned on egocentric video transfers well to third-person video understanding, robot learning, and even non-visual motion sequences such as inertial sensor data, suggesting that a single large-scale pre-training investment can amortize its carbon cost across a wide range of applications. Federated pre-training, in which models are trained collaboratively across many user devices without centralizing raw data, can further reduce the environmental impact of data transfer and cooling in massive datacenters, though it introduces communication overheads that must be carefully optimized. Metrics that combine model accuracy with CO2-equivalent emissions and device longevity should become standard reporting requirements alongside traditional performance benchmarks, incentivizing the community to treat energy proportionality as a first-class design objective.

Scalability also pertains to the sheer number of distinct activities and environments that a global egocentric recognition system must cover. A monolithic model trained to recognize every possible activity is likely to be unwieldy and brittle. The hierarchical design encourages a modular ecosystem where coarse-scale motion encoders are shared globally, while fine-scale decoders are specialized to particular domains, such as cooking, manufacturing, or clinical settings. This modularity allows for incremental updates to domain-specific components without retraining the entire stack, reducing both compute cost and downtime. Distributed versioning and model governance protocols must be established to manage the co-evolution of these modules, ensuring that an update to the fine-scale cooking stream does not inadvertently alter the behavior of the coarse-scale stream that other applications rely on. Lifecycle management becomes a socio-technical coordination challenge that demands clear

APIs, semantic versioning of motion representations, and ongoing monitoring for concept drift across the installed base.

8. Conclusion

Self-supervised hierarchical motion learning provides a conceptually elegant and architecturally powerful lens through which to address the enduring challenges of egocentric activity recognition in long videos. By structuring motion as a multiscale phenomenon that spans from fine-grained hand dynamics to extended daily routines, this framework aligns the inductive biases of the learning system with the natural organization of human behavior, thereby reducing annotation burden while enhancing temporal coherence and robustness. Our analysis has demonstrated that the benefits of such an approach cannot be realized through algorithmic innovation alone; they depend on deliberate choices across the full stack, including memory-compute co-design, training curriculum design, modality fusion, and conditional computation for energy efficiency. The hierarchical architecture creates new opportunities for privacy-aware edge deployment and modular system management, but it also introduces new attack surfaces, fairness vulnerabilities, and governance gaps that must be proactively addressed.

The path forward lies in embracing the deeply interdisciplinary nature of the problem. Computer vision and machine learning must enter sustained dialogue with human-computer interaction, law, ethics, and energy systems engineering. Pre-training datasets must be diversified through participatory methods, fairness metrics must become integral to the training workflow, and regulatory frameworks must evolve to recognize the distinctive risks of continuous behavioral inferencing. Hierarchical motion learning, when embedded within a thoughtfully designed socio-technical infrastructure, can empower assistive technologies, advance scientific understanding of daily activity, and support human autonomy. Without such embedding, it risks becoming yet another layer of opaque surveillance. The challenge, therefore, is not merely to build a more accurate recognizer, but to construct a learning system whose structure reflects the layered, situated, and ethically consequential nature of the activities it seeks to understand.

References

1. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. Proceedings of the 37th International Conference on Machine Learning (ICML), 1597–1607.

2. Grauman, K., Westbury, A., Byrne, E., et al. (2022). Ego4D: Around the world in 3,000 hours of egocentric video. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 18995–19012.

3. Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 6202–6211.

4. Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? Proceedings of the 38th International Conference on Machine Learning (ICML).

5. Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in Neural Information Processing Systems (NeurIPS), 35, 10078–10093.

6. Ryoo, M. S., Piergiovanni, A. J., Arnab, A., Dehghani, M., & Angelova, A. (2023). Hiera: A hierarchical vision transformer without the bells-and-whistles. Proceedings of the 40th International Conference on Machine Learning (ICML).

7. Yao, Y., Rosasco, L., & Caponnetto, A. (2022). On the robustness of self-supervised representations across domain shifts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12), 8844–8858.

8. Wu, C.-Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., & Girshick, R. (2019). Long-term feature banks for detailed video understanding. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 284–293.

9. Jin, Haopeng, et al. "HY-Himmel Technical Report: Hierarchical Interleaved Multi-stream Motion Encoding for Long Video Understanding." arXiv preprint arXiv:2605.08158 (2026).

10. Damen, D., Doughty, H., Farinella, G. M., et al. (2022). Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. International Journal of Computer Vision, 130(1), 33–55.

11. Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., & Gupta, A. (2016). Hollywood in homes: Crowdsourcing data collection for activity understanding. Proceedings of the European Conference on Computer Vision (ECCV), 510–526.

12. Pirsiavash, H., & Ramanan, D. (2012). Detecting activities of daily living in first-person camera views. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2847–2854.

13. Furnari, A., & Farinella, G. M. (2020). Rolling-unrolling LSTMs for action anticipation from first-person video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(11), 4021–4036.

14. Koppula, H. S., & Saxena, A. (2016). Anticipating human activities using object affordances for reactive robotic response. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(1), 14–29.

15. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Niebles, J. C. (2018). Dense-captioning events in videos. International Journal of Computer Vision, 126(2-4), 212–236.

16. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626.

17. Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT), 220–229.

18. Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4(2), 2053951717743530.

19. Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.

20. European Commission. (2021). Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM/2021/206 final.

21. Patterson, D., Gonzalez, J., Hölzle, U., et al. (2021). The carbon footprint of machine learning training will plateau, then shrink. Computer, 55(7), 18–28.

22. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 3645–3650.

23. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS).