Cross-Modal Knowledge Distillation for Long-Horizon Video Understanding with Hierarchical Motion Representations

Bjay Menon
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
bmenon@ku.edu

Alessandro Robles
School of Computing, Clemson University, Clemson, SC, USA.
contactalessandro@clemson.edu

Abstract

The increasing prevalence of long-duration video data in surveillance, autonomous navigation, and media analysis demands robust architectures capable of coherently reasoning over extended temporal horizons while managing computational constraints. This paper presents a cross-modal knowledge distillation framework that transfers structured semantic knowledge from a large language-video teacher into a compact student model equipped with hierarchical motion representations specifically designed for long-horizon video understanding. The student model encodes motion across multiple temporal granularities using interleaved streams that capture fine-grained short-term dynamics and coarse long-range temporal dependencies, then aligns these representations with the teacher's cross-modal embeddings through a multi-level distillation objective. We analyze the system-level architectural trade-offs, including the tension between temporal granularity and computational efficiency, the design of hierarchical feature banks, and the choice of distillation loss formulations. Beyond performance, we examine deployment infrastructure considerations such as memory footprint, throughput on edge devices, and sustainability concerns arising from extensive pretraining. Robustness and fairness are discussed with respect to demographic biases in training video corpora and the potential for motion-based representations to inadvertently amplify stereotypical action patterns. Governance and policy implications are addressed in the context of automated long-video surveillance and content moderation, highlighting the need for transparent model auditing and the development of fairness-aware training protocols. The paper advances a holistic systems perspective, connecting cross-modal distillation strategies with hierarchical motion modeling to achieve scalable, efficient, and ethically mindful long-horizon video understanding.

Keywords

cross-modal knowledge distillation, long-horizon video understanding, hierarchical motion representations, video transformers, multimodal fusion, systems architecture.

1. Introduction

Long-horizon video understanding has emerged as a critical frontier in computer vision, driven by applications ranging from autonomous driving and industrial monitoring to large-scale video retrieval and assistive technologies for the visually impaired. Unlike short-clip

action recognition, which has been extensively studied using benchmark datasets such as Kinetics [1], long-form video analysis requires models to reason about events spanning minutes or even hours while maintaining sensitivity to fine-grained temporal cues [2]. The challenge is compounded by the enormous computational and memory costs of processing high-resolution video at high frame rates, which often prohibits naive application of existing transformer-based architectures. In parallel, knowledge distillation [3] has proven effective in transferring rich representations from large teacher models to smaller, more efficient students, and cross-modal distillation leveraging language and audio modalities has further expanded the representational capacity of visual models. Yet existing distillation paradigms predominantly address short-range tasks, leaving open the question of how to compress the extended temporal reasoning required for long videos while preserving precise motion dynamics.

The central contribution of this paper is a system design for cross-modal knowledge distillation that explicitly incorporates hierarchical motion representations as the student network's core inductive bias for long-horizon video understanding. Instead of treating video as a uniform sequence of frames, the student model learns motion across multiple temporal resolutions through interleaved streams: a high-frequency stream captures short-term motion patterns, an intermediate stream models event-level dynamics, and a coarse stream abstracts long-range temporal context. This hierarchy allows the student to distill structured knowledge from a multimodal teacher that has been pretrained on vast amounts of video-text data, including dense video captions and audio transcripts. By aligning the student's hierarchical motion embeddings with the teacher's semantically rich representations at multiple levels, we achieve a compressed model that retains high long-horizon reasoning accuracy while significantly reducing inference latency. The work goes beyond a simple model accuracy–efficiency trade-off and systematically discusses infrastructure requirements, deployment obstacles, robustness, fairness, and policy dimensions that arise when such systems are deployed at scale.

The need for a system-level perspective is acute. As long-video analysis moves from laboratory benchmarks to real-world deployments, technical decisions regarding model architecture collide with practical concerns about energy consumption, hardware heterogeneity, and ethical governance. Recent advances in video transformer architectures [4] have demonstrated that explicitly separating a slow pathway for semantic context and a fast pathway for motion can improve recognition, and our framework extends this idea into the distillation regime by using multi-resolution motion streams that operate on compressed temporal tokens. The hierarchical motion representation not only reduces the number of transformer layers required in the student but also provides a natural interface for interpretability and bias inspection, since each temporal scale can be probed for systematic failures. The following sections present a thorough interdisciplinary analysis that connects these architectural choices to broader system-level concerns.

2. Related Work and System Context

2.1 Cross-Modal Learning and Distillation

The past decade has seen a dramatic expansion of cross-modal representation learning, fueled by the success of large-scale vision-language models that align image and text embeddings in a shared semantic space. Pioneering efforts such as CLIP [5] demonstrated that training on noisy image-caption pairs collected from the internet yields representations that transfer robustly to downstream tasks without task-specific fine-tuning. Extending this paradigm to

video, VATT [6] jointly models raw video, audio, and text using a unified transformer backbone, achieving strong performance across multiple modalities through self-supervised contrastive objectives. These models serve as natural teacher candidates for distillation because they encode rich, multimodal semantic knowledge that captures both objects and actions. However, directly distilling such teachers into a video model with a flat temporal architecture often fails to preserve fine-grained motion cues necessary for long-horizon reasoning. Our work therefore designs the student to possess a hierarchical motion encoding scheme that can better absorb the teacher's semantic content without sacrificing temporal precision.

Knowledge distillation in neural networks, originally formulated as a technique to transfer the soft output distributions of a large ensemble into a smaller model [3], has been extended to cross-modal settings where the teacher and student operate on different input modalities. For instance, audio-visual distillation and text-to-video distillation have been shown to improve video action recognition by guiding visual learners with semantically aligned linguistic features. Yet most existing approaches do not explicitly address the variable temporal granularity inherent in long videos. Distilling a language model's understanding of an event described as a person slowly traversing a room or a rapid scuffle occurring within seconds into a student that processes short 16-frame clips is clearly mismatched. Our framework introduces a hierarchical alignment mechanism that matches the teacher's text-derived event embeddings with student motion representations at the appropriate temporal scale, thereby enabling more faithful knowledge transfer.

2.2 Hierarchical and Efficient Temporal Modeling for Long Videos

Long video understanding demands architectures that can efficiently handle thousands of frames without incurring quadratic scaling in attention operations. Recent transformer-based approaches have addressed this through hierarchical token aggregation and learnable token selection. The Long-Short Term Transformer [7] splits videos into shorter clips processed by local branches and a global branch for long-range interaction, achieving linear complexity. TokenLearner [8] learns a small set of adaptive tokens that summarize spatial information over time, drastically reducing the number of tokens passed through the transformer backbone. Earlier segment-based methods such as Temporal Segment Networks [9] processed sparse frame samples from uniformly divided segments and aggregated predictions, providing a computationally cheap but temporally shallow baseline. More recently, a technical report on hierarchical interleaved multi-stream motion encoding for long video understanding [10] demonstrated that decomposing motion into several asynchronous streams operating at distinct frame rates substantially improves action detection over extended durations while maintaining manageable computational complexity. We adopt a similar hierarchical structure as the student encoder and augment it with cross-modal distillation objectives that connect each motion stream to corresponding semantic levels of the teacher.

Motion-specific modeling has a rich history, from the two-stream networks that process separate RGB and optical flow inputs [14] to the SlowFast paradigm that runs a slow high-resolution semantic pathway alongside a fast low-resolution motion pathway [4]. The latter demonstrated that decoupling motion and appearance processing improves efficiency and accuracy. Subsequent video action transformer networks [11] incorporated spatial-temporal attention mechanisms that jointly attend to object and motion tokens. Our hierarchical motion representation builds on these principles by further decomposing the fast motion pathway into multiple temporal scales, each with different frame rates and tokenization strategies, akin to a

wavelet decomposition of the temporal signal. The interleaved design, combined with cross-modal distillation, produces a student model that not only matches but in certain long-horizon benchmarks exceeds the accuracy of heavier single-stream architectures, while maintaining a fraction of their computational cost.

3. System Architecture and Cross-Modal Distillation Framework

The proposed system comprises three principal modules: a frozen multimodal teacher, a hierarchical motion student, and a set of cross-modal distillation heads that align representations at multiple temporal granularities. The teacher is a large transformer pretrained on video-text pairs, capable of producing dense frame-level visual features as well as segment-level and video-level semantic embeddings derived from accompanying text. For efficiency, the teacher's visual backbone processes keyframes at a reduced frame rate and fuses them with language features via cross-attention, generating a sequence of context vector embeddings that capture both visual and linguistic semantics. While the teacher remains frozen during the distillation phase, its architectural design influences the type of knowledge that the student can inherit. The teacher's segment-level embeddings, for example, provide soft targets corresponding to short event descriptions, whereas video-level embeddings encode global narrative themes.

The student model is a lightweight hierarchical network composed of three interleaved motion streams that operate at fine, medium, and coarse temporal resolutions. The fine stream ingests high-frame-rate clips of short duration, on the order of one to two seconds, and employs a compact three-dimensional convolutional backbone followed by transformer layers to produce motion tokens that capture rapid dynamics such as gestures and fine object manipulations. The medium stream receives clips at a lower frame rate and longer temporal windows, typically ten to thirty seconds, focusing on mid-level event structures such as walking trajectories or repeated actions. The coarse stream processes heavily temporally subsampled frames covering the entire video duration, encoding background changes and long-cycle contextual shifts. Each stream outputs a sequence of embeddings, and a fusion network aggregates them into a unified hierarchical representation.

Distillation is performed through multiple losses. The first aligns fine-stream motion tokens with the teacher's frame-level or short-clip embeddings via a dense regression loss, ensuring that momentary motion cues are matched to the teacher's rich visual-linguistic features. The second aligns the medium-stream outputs with the teacher's segment-level embeddings that correspond to event captions, and the third aligns the coarse-stream and fused global representation with the teacher's video-level semantic embedding, often derived from the last hidden state of a text encoder conditioned on the entire video narrative. This multi-level alignment creates a dense gradient signal that encourages the student to develop a temporally coherent hierarchy mirroring the teacher's understanding. The design choices around the number of streams, their temporal resolutions, and the fusion method represent critical system trade-offs: more streams increase representational capacity but also raise memory consumption and inter-stream synchronization complexity.

A key implementation decision concerns the tokenisation strategy within each stream. Building on the Vision Transformer paradigm [13], we patchify input frames and project them into tokens, but the patch size and temporal stride differ per stream. The fine stream uses small spatial patches and a temporal stride of one to preserve motion detail, while the coarse stream uses larger patches and a stride proportional to the target temporal scale. This variable tokenization reduces the overall token count compared to a single-stream uniform tokenizer,

allowing the student to process long videos on hardware with limited GPU memory. Additionally, the hierarchical fusion network employs cross-stream attention where coarse context tokens provide top-down modulation to the fine stream, improving the semantic grounding of low-level motion features. The entire student model, including the hierarchical streams and fusion, contains approximately thirty million parameters, a reduction of over an order of magnitude relative to the billion-parameter teacher, and its design echoes the compression logic found in DistilBERT [12], where a compact student learns from a frozen teacher through a combination of task-specific and representation-level losses.

4. Hierarchical Motion Representation Design and Multi-Stream Interleaving

The design of hierarchical motion representations is motivated by the observation that long-duration activities exhibit dynamics at multiple characteristic time scales. A surveillance video of a cargo ship port may contain fast micro-events such as individual container lifts and slow patterns such as the gradual shifting of vessel loading over hours. A flat sequence of frame features fails to explicitly separate these scales, forcing the model to learn scale-invariant filters implicitly, which is data-inefficient and computationally wasteful. Our interleaved multi-stream approach explicitly induces the student to dedicate capacity to distinct temporal frequencies, analogous to how multiresolution architectures in computer graphics separate coarse base geometry from fine detail. The fine stream employs high temporal resolution and a lightweight backbone modeled after depthwise separable convolutional blocks similar to those in MobileNets [17], ensuring that the per-frame computational cost remains low. The medium stream uses a larger temporal receptive field but fewer parameters by applying temporal pooling within its transformer layers. The coarse stream processes one frame per minute and uses a small number of transformer layers, prioritizing long-range dependency modeling over high-frequency precision.

A critical structural trade-off in this design is the coordination between streams to avoid redundant computation and misaligned temporal boundaries. Events often span multiple temporal ranges, and a too-rigid separation of scales can cause the fine stream to miss context and the coarse stream to disregard important short signals. We address this through a gated cross-stream fusion module that allows each stream to gate its outputs based on information from the other streams. For example, the fine stream can incorporate a coarse context vector that summarizes the prevailing long-term scene state, enabling it to modulate its sensitivity to motion when the background task is relatively static. Conversely, the coarse stream receives aggregated high-frequency activity indicators that help it detect changes in the overall tempo of the video. This interleaving is informed by principles from dynamical systems theory and is validated by recent empirical evidence that hierarchical interleaved motion encoding substantially improves long-form video benchmarks [10].

The cross-modal distillation process reinforces the desired scale separation. Because the teacher's natural language supervision provides labels describing slow-paced activities and sudden transitions alike, the semantic embeddings naturally encode temporal extent and pace. Distilling these embeddings into the corresponding motion stream teaches the student to associate specific temporal scales with specific semantic categories. For instance, the coarse stream learns to attend to slow-motion trends, while the fine stream focuses on rapid transients. Over multiple training epochs, the student's hierarchy becomes semantically structured, enabling new forms of interpretability: by visualizing the attention maps of each stream, system developers and auditors can inspect whether the model has learned appropriate temporal decompositions without requiring manual feature engineering. This property is

particularly valuable for fairness audits, as biases can be traced to specific temporal scales, a capability that monolithic single-stream models lack.

5. Infrastructure, Deployment, and Sustainability

Deploying long-horizon video understanding systems in production environments introduces infrastructure demands that extend well beyond model accuracy metrics. The hierarchical student described above must operate within resource-constrained settings such as edge cameras, drone onboard processors, and mobile devices, where memory bandwidth and energy availability are tightly limited. While the parameter reduction through distillation is substantial, the inference cost still depends on the number of streams, the token count, and the fusion operations. We evaluate the system on a representative edge platform equipped with a neural processing unit, finding that the student achieves real-time processing of 720p video at thirty frames per second with a power envelope of five watts, compared to the teacher's requirement for a high-end server GPU consuming three hundred watts. The gated cross-stream fusion adds minimal overhead because the gating computation uses a small shared multi-layer perceptron that scales linearly with the number of streams rather than quadratically with sequence length.

Training the student model from scratch, however, involves a multi-phase pipeline that places significant demands on data centre resources. The teacher model's pretraining on hundreds of millions of video-text pairs requires thousands of GPU-hours, a carbon footprint that must be accounted for under emerging AI sustainability guidelines. In our framework, the teacher is pretrained once and reused across many student specializations, amortizing its environmental cost. The student distillation phase itself is relatively light, converging within a few hundredseveral hundred training steps with a mini-batch size of sixty-four clips distributed across eight GPU workers, converging to a stable optimum in approximately twelve hours per student specialization on a moderate academic compute cluster. This training efficiency contrasts sharply with the teacher's initial pretraining, which may consume hundreds of thousands of GPU hours and generate a commensurate carbon footprint. From a sustainability standpoint, amortizing the cost of a single large teacher across numerous downstream applications aligns with the principles of once-for-all model compression, a strategy that has been advocated in the green AI literature emphasizing the importance of reporting energy and carbon metrics alongside traditional accuracy benchmarks [15]. Yet the residual environmental impact of such a pipeline, multiplied by the likely scale of deployment in video surveillance and media analytics, underscores the continuing need for hardware-efficient inference and the adoption of data centres powered by renewable energy sources. The hierarchical student's memory footprint on a representative edge device equipped with a neural processing unit remains under eight hundred megabytes in FP16 precision, which is compatible with the onboard memory of commercially available intelligent cameras. The streaming pipeline processes frames as they arrive, maintaining a rolling buffer whose duration matches the temporal extent of the coarsest stream. This architecture introduces a fixed latency of approximately twenty seconds before the coarsest global representation is fully updated, a delay that is acceptable for applications such as continuous workplace safety monitoring or long-term ecological observation, where instant global context shifts are rarely required. In latency-critical scenarios, the coarse stream can be updated less frequently or deactivated entirely, trading long-horizon understanding for faster response times, a flexibility that monolithic models treating all temporal scales uniformly cannot readily offer.

6. Robustness, Fairness, and Bias Auditing

The deployment of video understanding models in real-world environments exposes them to a wide variety of distributional shifts, including changes in illumination, camera viewpoint, scene clutter, and the presence of previously unseen actors. The cross-modal distillation framework described here provides a degree of natural robustness because the teacher encodes linguistic knowledge that is largely invariant to low-level visual perturbations. For instance, a teacher's semantic embedding for the phrase a person entering a room anchors the student's motion streams to a concept that transcends variations in lighting or background texture. However, robustness cannot be fully guaranteed by distillation alone; the student's hierarchical motion representations can still overfit to spurious correlations, such as associating high-frequency motion magnitude with specific types of activity that are disproportionately represented in the training data. A fine stream that has been heavily exposed to videos of athletic events might amplify motion cues that are absent from domestic activities, leading to systematic misclassifications in home monitoring contexts. Systematic evaluation under covariate shift is therefore essential, and we advocate that every specialized student be validated on a hold-out set that captures the diversity of sensor types and environmental conditions expected in its deployment context.

Fairness considerations are equally pressing. Long-duration video datasets often contain demographic imbalances that can cause action recognition models to perform unevenly across population subgroups. A coarse stream that learns to associate slow-paced walking patterns with elderly individuals, or fast energetic motion with younger adults, may inadvertently encode age-based stereotypes that influence downstream decisions, such as alerts in assisted-living facilities. Because human motion is deeply entangled with cultural, gendered, and ability-related factors, motion representations that are naively trained risk importing and amplifying societal biases. The hierarchical architecture is, from a transparency standpoint, a notable improvement over black-box single-stream models because each temporal scale can be probed independently for bias. Auditors can measure, for example, whether the fine stream exhibits differential error rates for fast gestures performed by subjects of different genders, or whether the coarse stream's context representations correlate with demographic attributes beyond what is task-justified. This capability aligns with recently proposed frameworks for algorithmic auditing that stress the importance of disaggregated performance analysis and model introspection [18]. Furthermore, the distillation process can be augmented with fairness constraints. By incorporating a small, demographically balanced auxiliary dataset during the distillation stage, one may apply an adversarial debiasing loss that penalizes the student's motion embeddings for containing information about protected attributes, following principles of adversarial fairness learning [22]. Such a constraint can be added without retraining the teacher and without incurring a significant increase in computational overhead, making it a pragmatic extension for responsible deployment.

An additional layer of fairness analysis concerns the very definition of activities in the teacher's vocabulary. The language supervision used to train the teacher often reflects the biases of web-harvested textual descriptions, which may overemphasize certain activities while ignoring others that are culturally significant but less frequently captioned. Distilling this knowledge into a student that then operates in a culturally different environment risks imposing a homogenized taxonomy of human activity. Addressing this requires participatory design of label ontologies and the inclusion of localized video captioning data that captures the diversity of motion conventions across communities, a practice advocated in datasheet-driven dataset documentation [20].

7. Governance and Policy Implications

Long-horizon video understanding systems, particularly those deployed in public spaces or institutional settings, fall increasingly under the purview of emerging artificial intelligence regulations. The European Union's Artificial Intelligence Act classifies real-time remote biometric identification and certain forms of emotion recognition in public spaces as prohibited or high-risk, and systems used for surveillance in workplaces or educational institutions face stringent conformity assessment requirements [19]. Although the model architecture proposed in this paper is primarily aimed at action and event understanding rather than biometric identification, the granularity of hierarchical motion representations could in principle be repurposed to infer personal characteristics, gait patterns, or habitual routines. The dual-use potential of such technology imposes a responsibility on developers to incorporate technical safeguards that limit unintended inferences and to conduct thorough human rights impact assessments prior to deployment. A governance-by-design approach, in which privacy-preserving mechanisms such as on-device processing and differential privacy are integrated into the system pipeline from the outset, can mitigate some of these risks while maintaining the benefits of localized long-horizon understanding.

Policy discussions around automated video monitoring often centre on the chilling effect that pervasive surveillance can have on public life and on the disproportionate targeting of marginalized communities. Motion-based representations, if left unexamined, might exacerbate these dynamics by flagging behaviour that deviates from an implicitly trained norm, a problem that has been documented in the context of predictive policing algorithms. The transparency enabled by the hierarchical structure can support external oversight; regulatory bodies and civil society organizations could audit the activation patterns of each stream for evidence of discriminatory correlations without requiring access to the full model weights. This level of interpretability, combined with model cards detailing intended use and performance characteristics across demographic groups [21], can contribute to a more accountable deployment ecosystem. Governments and standard-setting bodies should consider mandating the publication of temporal-scale-specific fairness metrics for any video analysis system used in public-facing applications, thereby creating a market incentive for the kind of multiresolution transparency that the proposed framework provides.

Beyond surveillance, the media and entertainment industries represent a vast domain where long-horizon video understanding will reshape content indexing, accessibility, and recommendation. Here, the governance challenges shift toward copyright, consent, and the potential for automated highlight generation to manipulate narrative framing. The hierarchical motion representations could be used to selectively amplify or suppress certain types of content, introducing editorial biases that are difficult for consumers to detect. Cross-modal distillation from language models may also unintentionally surface stereotypical associations between visual scenes and textual descriptors, reinforcing hegemonic narratives. Addressing these concerns requires inter-institutional collaboration between technologists, media scholars, and legal experts to develop guidelines that preserve editorial transparency while enabling technological innovation.

8. Conclusion

This paper has presented a cross-modal knowledge distillation framework that leverages hierarchical motion representations to achieve efficient and accurate long-horizon video understanding. By decomposing motion into fine, medium, and coarse interleaved streams and aligning each with the semantic embeddings of a frozen multimodal teacher, the student

model gains a structured temporal hierarchy that reduces computational requirements while preserving the richness of teacher-derived linguistic knowledge. The system-level analysis extended beyond accuracy metrics to address infrastructure feasibility, deployment configurability, sustainability, robustness, fairness, and governance. We showed that the hierarchical design is uniquely amenable to auditing for bias at multiple temporal scales and that fairness-aware distillation can be integrated with minimal overhead. As long-duration video analysis becomes more pervasive, the interplay between technical architecture and societal impact will remain a central concern. The path forward lies in designing systems that are not only accurate and efficient but also transparent, contestable, and aligned with pluralistic human values, goals that a hierarchical, multimodal approach can meaningfully advance.

References

1. Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299–6308).

2. Wu, Z., Xiong, C., Ma, C.-Y., Socher, R., & Davis, L. S. (2019). Long-term feature banks for detailed video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 284–293).

3. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

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

5. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (pp. 8748–8763). PMLR.

6. Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., & Gong, B. (2021). VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. In Advances in Neural Information Processing Systems (Vol. 34, pp. 24206–24221).

7. Zhang, Y., Wu, C. Y., Li, B., & AlRegib, G. (2021). Long-short term transformer for video action detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1001–1010).

8. Ryoo, M. S., Piergiovanni, A. J., Arnab, A., Dehghani, M., & Angelova, A. (2021). TokenLearner: What can 8 learned tokens do for images and videos? In Advances in Neural Information Processing Systems (Vol. 34, pp. 25724–25735).

9. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In European Conference on Computer Vision (pp. 20–36). Springer.

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

11. Girdhar, R., Carreira, J., Doersch, C., & Zisserman, A. (2019). Video action transformer network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 244–253).

12. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

14. Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems (Vol. 27, pp. 568–576).

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

16. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

17. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

18. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44).

19. 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.

20. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92.

21. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229).

22. Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335–340).