

Transformer-Based Modeling of Multimodal Physiological Signals for Human Activity Recognition

Fanhaoran Zhu

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
zhu1995@buffalo.edu

Sonia Upadhyay

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis,
OR, USA.
soniau247@oregonstate.edu

Nathan R. Howard

Department of Computer Science, University of Central Florida, Orlando, FL, USA.
contactnathan@ucf.edu

Abstract

Human activity recognition has long relied on hand-crafted features and relatively shallow classifiers applied to individual sensing modalities. The rise of deep learning created new possibilities for learning directly from time-series data, but many architectures still struggle to capture long-range dependencies, multimodal interactions, and heterogeneous sampling regimes found in physiological sensing. Transformer-based models offer a promising alternative because they can attend flexibly across temporal positions and sensor channels without being constrained by strict Markov assumptions or fixed receptive fields. This paper presents a system-level analysis of transformer-based modeling for multimodal physiological signals in human activity recognition. It examines sensing infrastructure, data governance, architectural trade-offs, training strategies, deployment constraints, robustness, fairness, and sustainability. Rather than focusing on a single algorithmic contribution, the paper discusses how multimodal physiological data from wearable, optical, electrodermal, cardiac, and inertial sensors can be organized and processed through attention mechanisms. The analysis includes structural trade-offs involving centralized and distributed learning, memory and energy budgets, cross-modal fusion order, interpretability, and regulatory compliance. Special attention is given to fairness across populations, generalization under sensor variability, and the governance of health-related inference systems. The paper argues that transformer-based activity recognition should be treated not only as a modeling problem but as a socio-technical infrastructure challenge requiring careful balancing of accuracy, resource consumption, privacy, and accountability. A forward-looking perspective is offered on the deployment of these systems in clinical, industrial, and everyday contexts.

Keywords

human activity recognition, multimodal physiological signals, transformer architectures, attention mechanisms, wearable systems, system governance, robustness, fairness, sustainability.

1. Introduction

Human activity recognition has become a foundational capability in pervasive computing, clinical monitoring, industrial safety, and consumer health applications. The field has evolved from rule-based classifiers and carefully engineered features toward end-to-end deep learning pipelines that learn representations directly from sensor streams. The introduction of transformer architectures reshaped sequence modeling by replacing recurrent or convolutional inductive biases with scalable attention mechanisms that can capture interactions across long temporal distances [1]. This shift has influenced natural language processing, where pretrained bidirectional models established new performance standards [2], and then extended into computer vision through patch-based image representations [3]. In time-series analysis, transformer-based approaches have been increasingly adapted to multivariate physiological and motion data because they can model complex dependencies that are difficult to encode with fixed temporal windows or recurrent state updates [4]. These developments do not eliminate the relevance of earlier neural architectures. Long short-term memory networks and convolutional models remain useful baselines and components, particularly when computational constraints favor local operations and recurrent gating [5]. The broader transformation of machine learning from feature engineering to representation learning has nonetheless created new opportunities for activity recognition systems that can ingest heterogeneous signals and adapt across tasks [6].

Wearable and ambient sensing systems now routinely collect multiple synchronous physiological streams, including electrocardiography, photoplethysmography, electrodermal activity, skin temperature, accelerometry, gyroscopy, and occasionally optical or biochemical signals. Human activity recognition using body-worn inertial sensors has been extensively studied and continues to inform system design [7]. Deep convolutional and recurrent architectures have been applied to multimodal wearable data to improve recognition accuracy in real-world settings [8]. Public benchmarks have supported reproducible comparisons across activity classes and sensor placements [9]. Surveys of deep learning for sensor-based activity recognition indicate that data heterogeneity, class imbalance, and cross-domain generalization remain persistent challenges even as model capacity increases [10]. The present paper extends this discussion by treating transformer-based multimodal physiological activity recognition as a systems problem. The focus is not on a single architecture or dataset but on the structural, infrastructural, and governance dimensions that determine whether such systems can be deployed responsibly and sustainably.

2. Background and System Context

Transformer-based time-series modeling has received substantial attention because it offers a flexible mechanism for capturing relationships among irregularly spaced and multimodal observations. Surveys of transformer models in time series highlight their ability to handle multivariate sequences, long context, and pretraining strategies originally developed for language or vision [11]. General surveys of transformer architectures have documented their rapid diffusion across domains and their recurring bottlenecks, including quadratic attention complexity, large memory requirements, and sensitivity to training data scale [12]. For physiological activity recognition, these considerations are intensified by the need to process privacy-sensitive biosignals in environments that may be constrained by battery life, connectivity, and regulatory oversight. The system context therefore includes not only model accuracy but also the conditions under which data are collected, stored, transmitted, and used for inference.

A significant system-level concern is the social and demographic representativeness of training data. Physiological signals vary across age, sex, skin tone, fitness level, medication use, and health status. When these factors are not explicitly accounted for, models may learn spurious associations or perform unevenly across groups. Bias and fairness research has shown that machine learning systems often inherit and amplify disparities from data collection and labeling processes [13]. In activity recognition, uneven sensor placement, motion artifacts, and differences in skin perfusion can produce modality-specific failures that are difficult to detect using aggregate accuracy alone. Transformer models, because they are highly expressive, may exploit confounding correlations across modalities unless fairness and robustness are designed into the data and evaluation pipeline. These socio-technical considerations are inseparable from the architectural and infrastructure decisions discussed in later sections.

3. Multimodal Physiological Signal Infrastructure

A transformer-based activity recognition system depends on a well-designed sensing and data infrastructure before any training occurs. The infrastructure must support synchronized acquisition of multiple modalities, each with its own sampling rate, noise profile, dynamic range, and failure mode. Inertial sensors often operate at tens or hundreds of hertz, while electrodermal activity and temperature signals change more slowly. Cardiac signals may require higher temporal resolution to preserve waveform morphology, whereas optical signals can be affected by ambient light and motion artifacts. These differences create substantial engineering challenges for alignment, buffering, timestamping, and quality control. If the infrastructure does not preserve precise temporal correspondence, attention mechanisms may learn relationships that reflect artifact structure rather than true physiological coupling.

The placement of processing logic across edge devices, gateways, and cloud resources is another structural decision. On-device processing reduces privacy exposure and latency but imposes limits on memory, energy, and model complexity. Cloud processing enables larger transformer models and centralized training but creates dependencies on connectivity and raises data protection concerns. A hybrid architecture may perform lightweight preprocessing and feature extraction at the edge while reserving heavier attention-based inference for a gateway or cloud service. This trade-off affects not only technical performance but also user acceptance, regulatory compliance, and operational resilience. In health-related applications, the ability to maintain functionality during connectivity loss or sensor dropout is especially important. Infrastructure design should therefore treat data availability and model inference as part of a single reliability architecture rather than as separate concerns.

Multimodal physiological data also raise questions of ownership, consent, and secondary use. Continuous biosignal collection may reveal information beyond physical activity, including emotional state, cardiovascular function, or metabolic condition. The infrastructure must therefore include access controls, audit logs, and mechanisms for data minimization. Missing this dimension can undermine trust even when the model itself performs well. From a systems perspective, the data pipeline is a governance artifact as much as a technical artifact. Activity recognition research has historically prioritized classification accuracy, but real-world deployments require that accuracy be achieved within an acceptable privacy and accountability envelope. Transformer models, with their capacity to memorize fine-grained patterns, increase the importance of these safeguards because they may expose sensitive information through model inversion or membership inference if not properly controlled.

4. Transformer-Based Modeling Architecture

The core architectural idea in transformer-based activity recognition is to represent each timestamp or sensor segment as an embedding and then allow attention operations to reweight these representations according to learned relevance patterns. This design removes the strict locality of convolutional approaches and the sequential bottleneck of recurrent models. It enables the model to relate a cardiac event observed seconds earlier to a movement pattern occurring in the present, provided that sufficient positional or temporal encoding is preserved. The original transformer formulation demonstrated the power of scaled dot-product attention for sequence transduction [1]. Vision transformers showed that the same principle can be applied to nonlinguistic inputs by dividing data into patches [3]. Time-series transformer methods extend this idea by treating time windows or channel segments as tokens that can be attended to jointly [4]. For multimodal physiological signals, the tokenization strategy can be based on fixed time windows, event-triggered segments, or learned projections from raw sensor streams.

A central architectural trade-off is the point at which modalities are fused. Early fusion concatenates raw or lightly processed signals before attention, allowing the model to discover cross-modal interactions from the start. This approach can capture fine-grained physiological relationships but may be sensitive to differing noise levels and sampling rates. Late fusion processes each modality independently and combines representations before classification, which improves modularity and interpretability but may miss interactions that require joint temporal reasoning. Hybrid fusion introduces cross-modal attention at intermediate layers, allowing one modality to query information from another. A recent study on joint attention mechanism learning for opto-physiological monitoring during physical activity illustrates how attention can be used to coordinate optical and physiological signals under motion artifacts [14]. This line of work demonstrates that cross-modal attention is not merely a modeling convenience but can serve as a mechanism for learning which sensors should be trusted under changing conditions. In system terms, the fusion architecture determines how much processing can be distributed, how easily the model can be updated, and how failures in one modality propagate to the final decision.

Transformer architectures for physiological data also raise important computational concerns. Attention over long sequences can become memory-intensive, especially when multiple modalities are tokenized at high frequency. Hierarchical attention, sparse attention, and segment-level recurrence are common structural responses that reduce complexity while preserving long-range context. These choices affect deployment feasibility on wearable devices and edge nodes. Another architectural concern is the use of pretraining. Self-supervised objectives such as masked signal prediction or contrastive learning across modalities can produce representations that transfer to downstream activity classes, but they require large and diverse datasets. Without pretraining, transformer models may underperform recurrent or convolutional baselines on small datasets. The system designer must therefore evaluate whether the additional capacity of attention-based models justifies the associated data, compute, and validation costs. In many practical settings, a hybrid architecture that uses lightweight convolutional front ends followed by selected attention layers may offer the best balance between accuracy and resource use.

5. Training and Evaluation Trade-offs

Training transformer-based multimodal activity recognition systems involves several decisions beyond hyperparameter selection. Data alignment, class definition, label quality, and evaluation protocol shape the meaning of reported performance. In health-related contexts,

responsible machine learning frameworks emphasize that high aggregate accuracy is insufficient without evidence of clinical validity, safety, and robustness across relevant subpopulations [15]. This is especially true when activity recognition is used to guide rehabilitation, monitor chronic conditions, or assess functional capacity. The training pipeline must therefore include uncertainty estimation, calibration, and careful handling of ambiguous transition periods between activities. Activity boundaries are often fuzzy, and rigid frame-level labels can mislead both training and evaluation.

Recurrent and convolutional baselines remain valuable points of comparison because they are often more computationally predictable and easier to certify in constrained environments [5][18]. Deep recurrent architectures have demonstrated strong performance for wearable activity recognition, particularly when data volumes are moderate and temporal dependencies are relatively local [18]. Smartphone-based studies using a single accelerometer have shown that even compact sensing platforms can support useful activity classification when models are matched to the data distribution [19]. Transformer-based systems should be evaluated against such baselines under identical data splits and deployment constraints, rather than only against other large models. This comparative discipline helps reveal whether attention mechanisms provide meaningful gains or merely add complexity.

Evaluation protocols should also account for sensor dropout, modality loss, and temporal misalignment. A robust multimodal system should degrade gracefully when one signal becomes unavailable. This can be tested by systematically masking modalities during inference and measuring the impact on activity recognition. In addition, evaluation should distinguish between user-dependent and user-independent settings. User-dependent protocols may overestimate real-world performance because the model has seen similar individuals during training. Cross-user, cross-device, and cross-environment protocols provide stronger evidence for generalization. These evaluation choices have direct consequences for regulatory claims and for the fair distribution of system benefits. A model that performs well only on the training population may create disparities when deployed in more diverse settings.

6. Robustness, Fairness, and Generalization

Physiological activity recognition systems are particularly vulnerable to distribution shift because sensor characteristics, human physiology, and environmental conditions vary widely. Robustness requires more than data augmentation; it requires an architectural and evaluation strategy that anticipates plausible failures. Transformer models may learn highly nonlinear decision boundaries that are sensitive to small perturbations in input signals. This is a concern for safety-critical applications where motion artifacts or sensor degradation should not lead to confident misclassification. Fairness is closely related to robustness. Bias can enter through nonrepresentative training data, inconsistent labeling practices, or differences in sensor performance across demographic groups [13]. Sociotechnical perspectives on fairness caution that abstract statistical parity metrics may fail to capture the institutional and contextual dimensions of harm [16]. In activity recognition, fairness should be assessed in terms of whether different populations receive comparable error rates for clinically meaningful activities, and whether errors lead to unequal consequences.

Interpretability is another system requirement. Attention weights are often presented as explanations, but their reliability as causal accounts of model behavior is contested. Governance frameworks increasingly require that algorithmic systems be auditable and that developers provide meaningful explanations for decisions that affect individuals [17]. For transformer-based multimodal models, explaining a prediction may require tracing how

different modalities influence the final attention distribution and how uncertainty accumulates across time. This is significantly harder than explaining a decision tree or a linear model. System designers should therefore pair attention models with post hoc explanation tools, uncertainty indicators, and human review pathways where decisions have high stakes. Accountability mechanisms are not external additions; they shape the data collection, model training, and deployment choices made throughout the lifecycle.

7. Deployment and Governance Considerations

Deployment of transformer-based multimodal activity recognition systems requires careful attention to computational infrastructure and organizational governance. Federated learning offers one path for training models across distributed devices without centralizing sensitive physiological data [20]. In federated settings, each device or site updates a local model and shares only model parameters or gradients with a coordinating server. This reduces direct data exposure but introduces new challenges, including communication costs, non-identically distributed data, and the risk of model parameter leakage. Transformer models with large parameter counts can be difficult to train under federated constraints because uplink bandwidth and on-device memory may be limited. A practical deployment may therefore combine federated pretraining on local devices with centralized fine-tuning on de-identified or synthetic data. The system architecture must specify how model updates are validated, how model versions are tracked, and how users can withdraw consent.

Governance also includes regulatory compliance in health and workplace settings. Activity recognition systems that infer health-related conditions may fall under medical device rules, data protection laws, or occupational safety standards. This creates obligations for documentation, risk management, and post-market surveillance. Internal algorithmic audit frameworks provide a way to integrate these obligations into the development process [17]. Audits should examine training data provenance, model performance across subgroups, failure modes, and the processes for contesting automated decisions. Without such mechanisms, technical teams may optimize for narrow accuracy metrics while external stakeholders bear the consequences of errors. The governance structure must also define who is responsible when a model misses a fall, misclassifies a high-risk activity, or produces an unfair outcome for a particular worker or patient.

8. Sustainability and Policy Implications

Sustainability in transformer-based activity recognition involves energy consumption, hardware lifecycle, and long-term maintenance of data and model infrastructure. Large transformer models require substantial computational resources during training, and repeated retraining or hyperparameter search can multiply environmental costs. In edge deployments, the energy budget of continuous sensing and on-device inference affects battery life and device longevity. Attention mechanisms that improve accuracy may not be justified if they require frequent recharging or replacement of wearable devices, particularly in low-resource settings. System-level sustainability analysis should therefore include not only training-time carbon estimates but also inference-time energy use, data transmission patterns, and the durability of sensing hardware.

Policy implications arise from the intersection of health data, algorithmic decision-making, and public infrastructure. Standards for multimodal physiological data interoperability can reduce fragmentation and enable more robust validation, but they must be designed with privacy and consent protections. Policies that encourage open benchmarks and transparent

evaluation can improve scientific reproducibility, yet they also risk concentrating power in organizations that control large datasets. A balanced policy approach would support federated research networks, require public reporting of subgroup performance, and establish clear rules for the secondary use of biosignal data. Sustainability policies should encourage model compression, selective inference, and hardware-aware training rather than treating accuracy as the sole objective. As transformer-based systems move from research prototypes into health care, occupational safety, and consumer platforms, their design choices will shape not only technical performance but also the distribution of benefits and risks across society.

9. Conclusion

Transformer-based modeling of multimodal physiological signals for human activity recognition represents a significant architectural shift from earlier recurrent and convolutional pipelines. Attention mechanisms offer the flexibility to integrate heterogeneous sensor streams, capture long-range temporal dependencies, and adapt across tasks. However, the system-level requirements for these models extend well beyond classification accuracy. Data infrastructure must support synchronization, privacy, and consent across modalities with very different physical characteristics. Architectural choices about fusion order, tokenization, and attention complexity determine feasibility on resource-constrained devices. Training and evaluation protocols must reflect realistic deployment conditions, including missing modalities, population diversity, and ambiguous activity boundaries. Robustness, fairness, and interpretability are not secondary concerns but fundamental design constraints that shape whether these systems can be trusted. Governance and policy frameworks must address accountability, auditability, and regulatory compliance, especially when physiological signals reveal health-related information. Sustainability considerations further require balancing model capacity against energy use, hardware lifecycle, and long-term maintenance. A successful transformer-based activity recognition system will therefore be one that integrates technical innovation with careful attention to infrastructure, equity, and institutional responsibility.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *Proceedings of the 9th International Conference on Learning Representations*.
4. Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., & Eickhoff, C. (2021). A transformer-based framework for multivariate time series representation learning. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2114–2124.

5. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
6. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
7. Bulling, A., Blanke, U., & Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys*, 46(3), 1–33.
8. Ordóñez, F. J., & Roggen, D. (2016). Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1), 115.
9. Reiss, A., & Stricker, D. (2012). Introducing a new benchmarked dataset for activity recognition. *Proceedings of the 16th International Symposium on Wearable Computers*, 108–109.
10. Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119, 3–11.
11. Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2023). Transformers in time series: A survey. *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 6778–6786.
12. Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, 3, 111–132.
13. 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.
14. Zheng, X., Hu, S., Dwyer, V., Barrett, L., & Derakhshani, M. (2026). Joint attention mechanism learning to facilitate opto-physiological monitoring during physical activity. *Biomedical Signal Processing and Control*, 113, 108949.
15. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), 1337–1340.
16. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68.
17. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44.
18. Murad, A., & Pyun, J.-Y. (2017). Deep recurrent neural networks for human activity recognition. *Sensors*, 17(11), 2556.
19. Chen, Y., & Xue, Y. (2015). A deep learning approach to human activity recognition based on single accelerometer. *Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics*, 1488–1492.
20. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P.

B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., & Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210.